FYP Final Presentation

S2TPVFormer: Improving 3D Semantic Occupancy Prediction using Spatiotemporal Transformers

Group 05

E/17/331 H.S.C. Silva
E/17/369 S.B. Wannigama

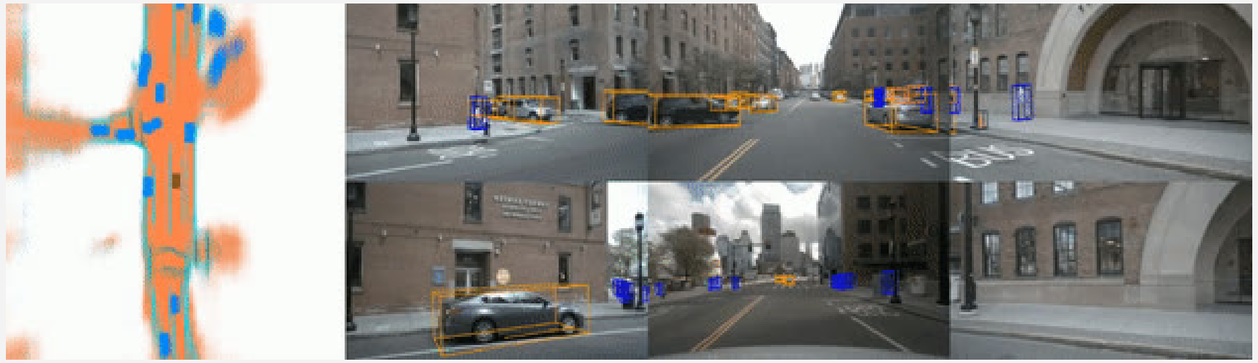
Supervisors

Prof. Roshan Ragel
Gihan Jayatilaka

3D Semantic Occupancy Prediction

Introduction

- Comprehensive **3D scene understanding & reasoning** are essential for building the perception stack of autonomous driving & robotic systems
- Similar Tasks: 3D Detection, Tracking Map Segmentation, Optical-Flow Estimation, **SOP**, etc.



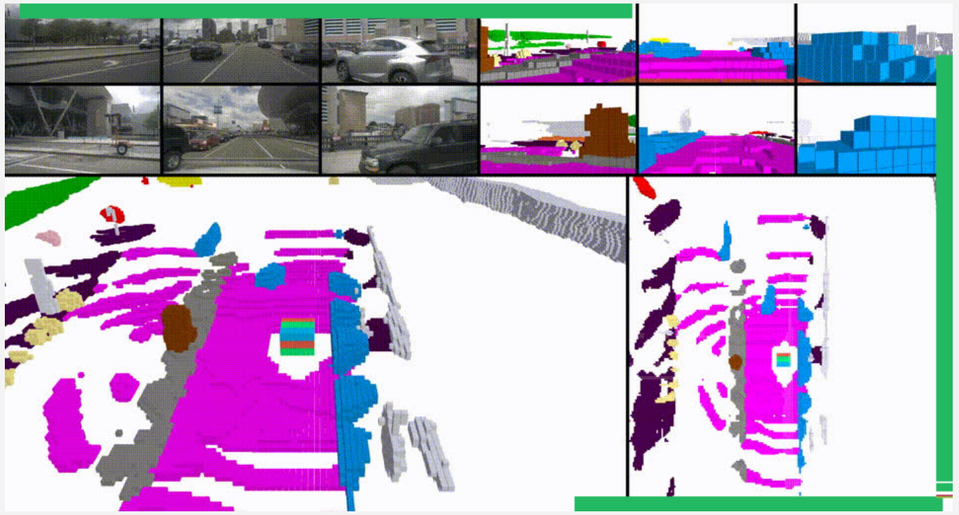


Figure 1: Comparison of Different approaches for 3D Scene Understanding

Figure 1 (left) shows visualizations of **Map Segmentation** and **3D Detection** tasks by Zhiqi Li et al.: BEVFormer; Figure 1 (right) shows a visualization of **3D SOP** task by Yuanhui Huang et al.: TPVFormer [2]

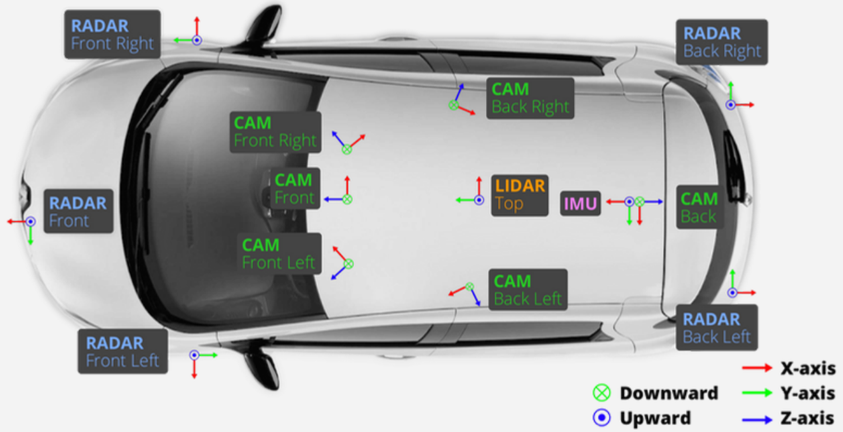


Figure 2: Different sensors available in a self-driving car

- Camera perception is better!
  - Cost-effective, Can detect long-range distance objects, Can identify road elements (ex: traffic lights, road signs, etc.)
- SOP: Predict semantics for all voxels in a given range, given a set of surrounding camera perspective views

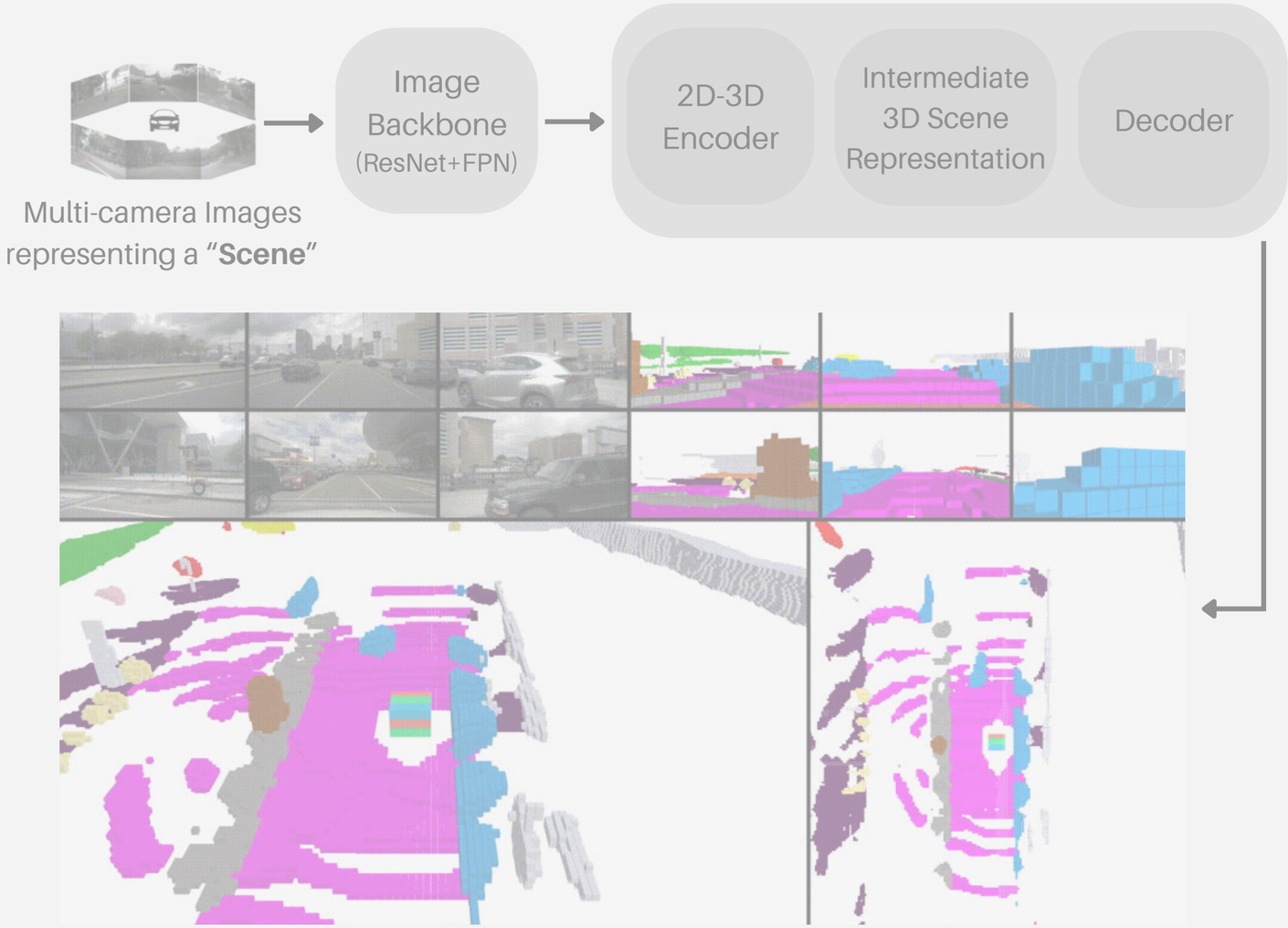


Figure 3: 3D Semantic Occupancy Prediction General Workflow [3]

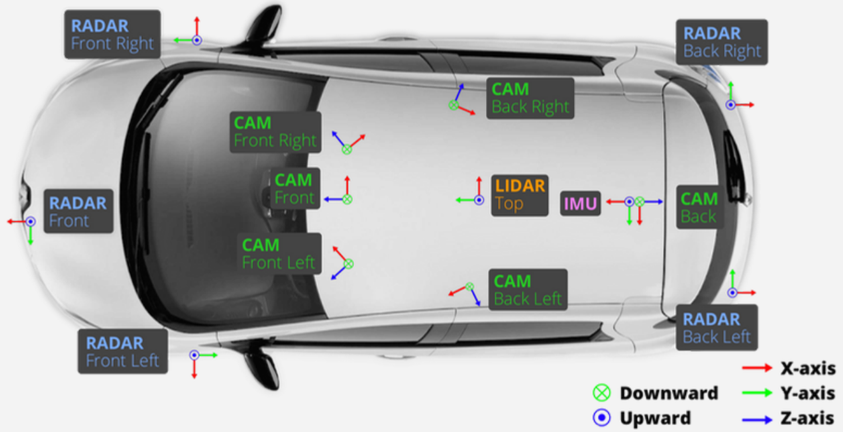


Figure 2: Different sensors available in a self-driving car

- Camera perception is better!
  - Cost-effective, Can detect long-range distance objects, Can identify road elements (ex: traffic lights, road signs, etc.)
- SOP: Predict semantics for all voxels in a given range, given a set of surrounding camera perspective views

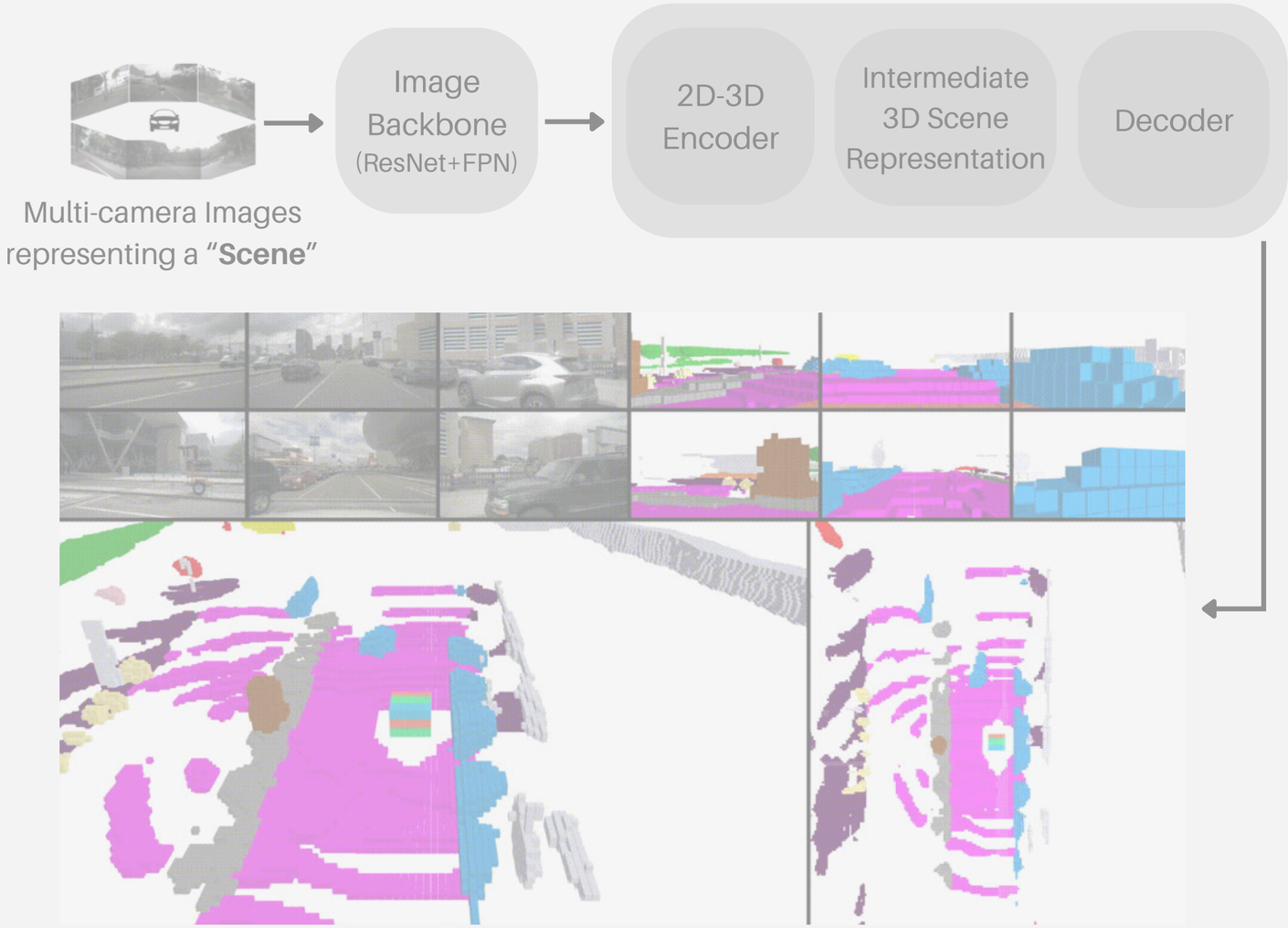


Figure 3: 3D Semantic Occupancy Prediction General Workflow [3]

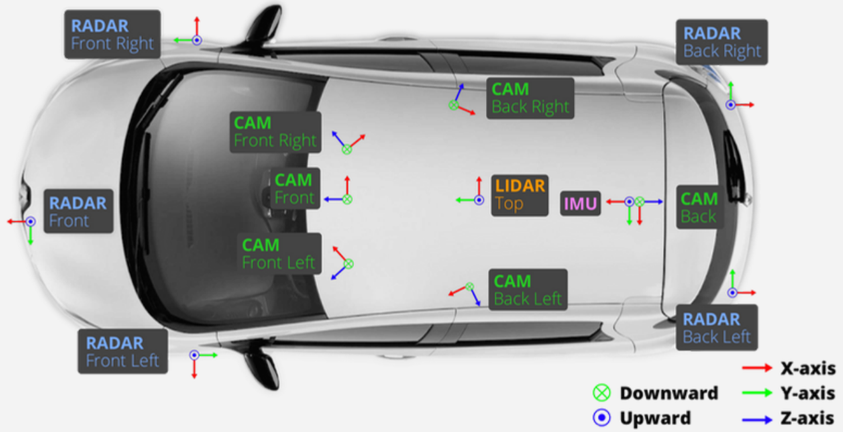


Figure 2: Different sensors available in a self-driving car

- Camera perception is better!
  - Cost-effective, Can detect long-range distance objects, Can identify road elements (ex: traffic lights, road signs, etc.)
- SOP: Predict semantics for all voxels in a given range, given a set of surrounding camera perspective views

Multi-camera Images representing a "**Scene**"

Image Backbone (ResNet+FPN)

2D-3D Encoder

Intermediate 3D Scene Representation

Decoder

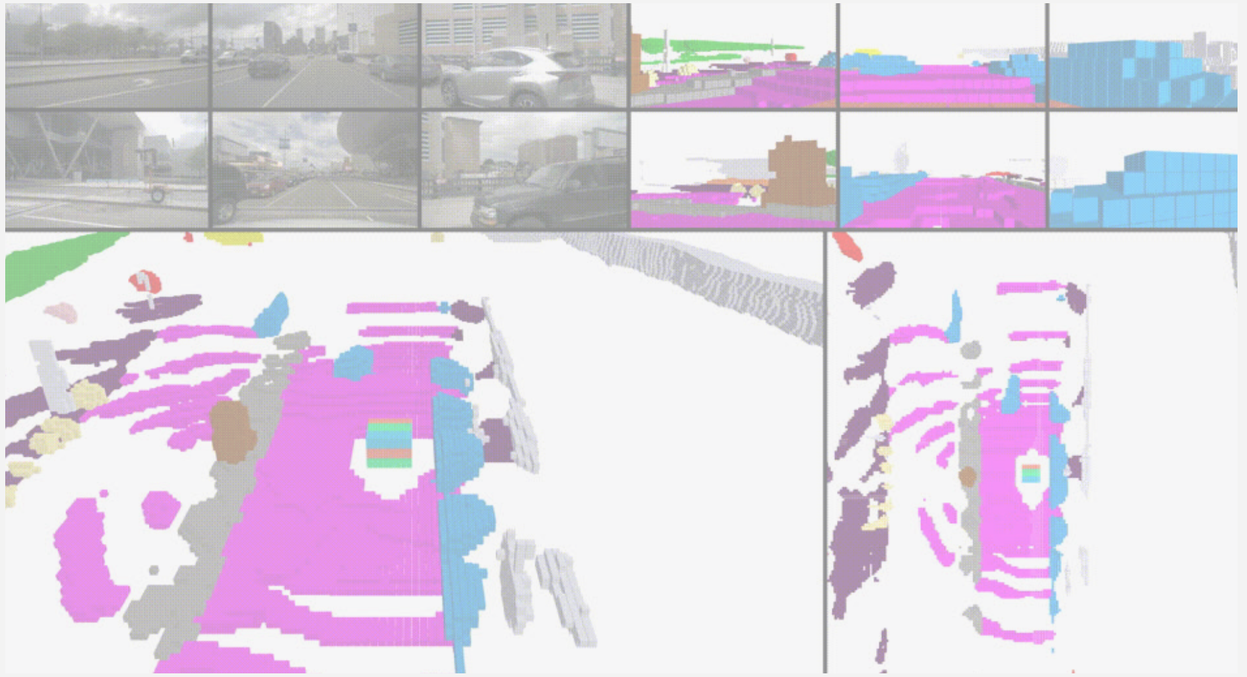


Figure 3: 3D Semantic Occupancy Prediction General Workflow [3]

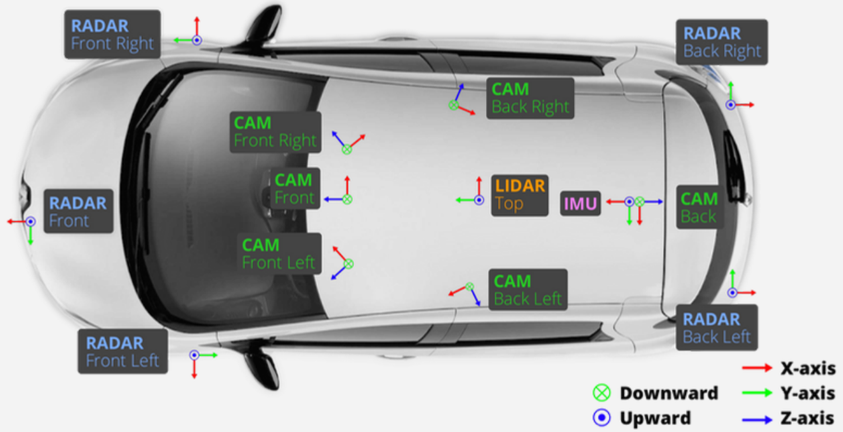


Figure 2: Different sensors available in a self-driving car

- Camera perception is better!
  - Cost-effective, Can detect long-range distance objects, Can identify road elements (ex: traffic lights, road signs, etc.)
- SOP: Predict semantics for all voxels in a given range, given a set of surrounding camera perspective views

Multi-camera Images representing a "**Scene**"

Image Backbone (ResNet+FPN)

2D-3D Encoder

Intermediate 3D Scene Representation

Decoder

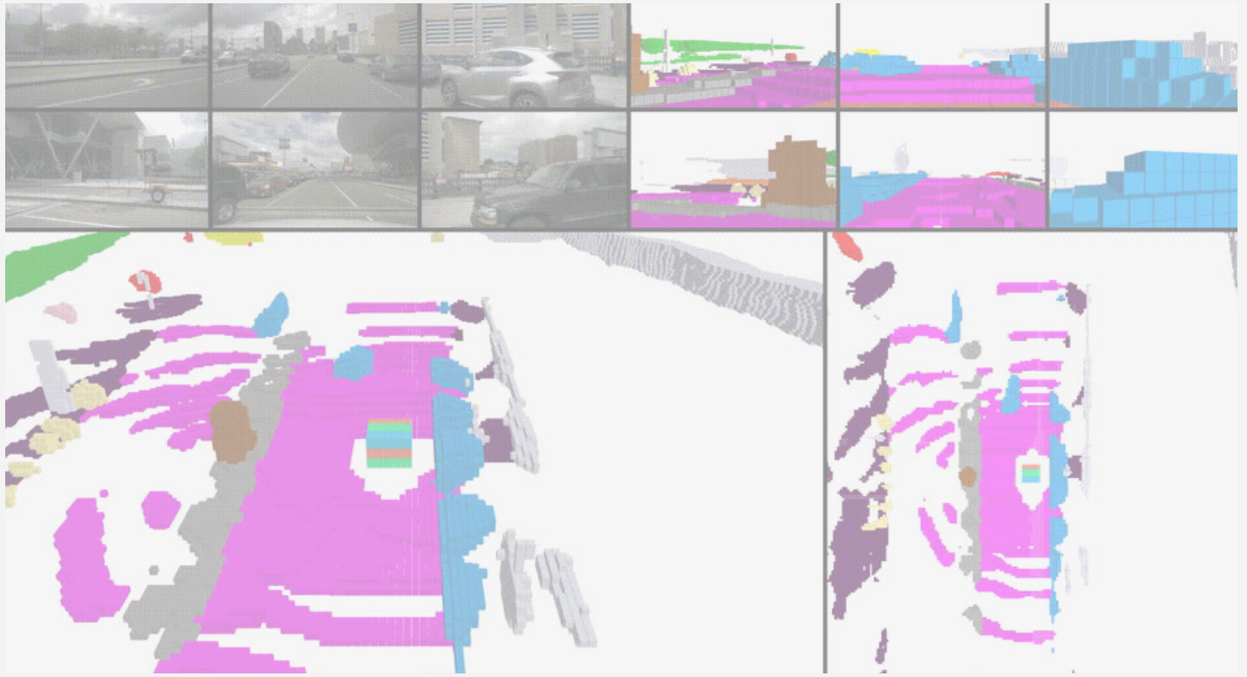


Figure 3: 3D Semantic Occupancy Prediction General Workflow [3]

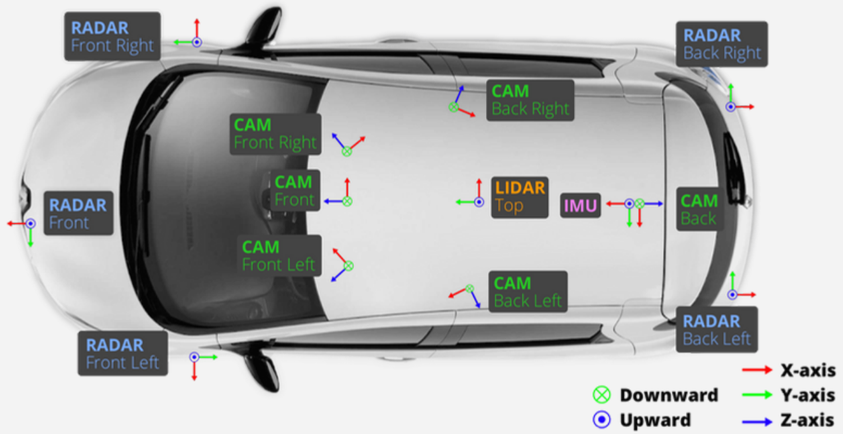


Figure 2: Different sensors available in a self-driving car

- Camera perception is better!
  - Cost-effective, Can detect long-range distance objects, Can identify road elements (ex: traffic lights, road signs, etc.)
- SOP: Predict semantics for all voxels in a given range, given a set of surrounding camera perspective views

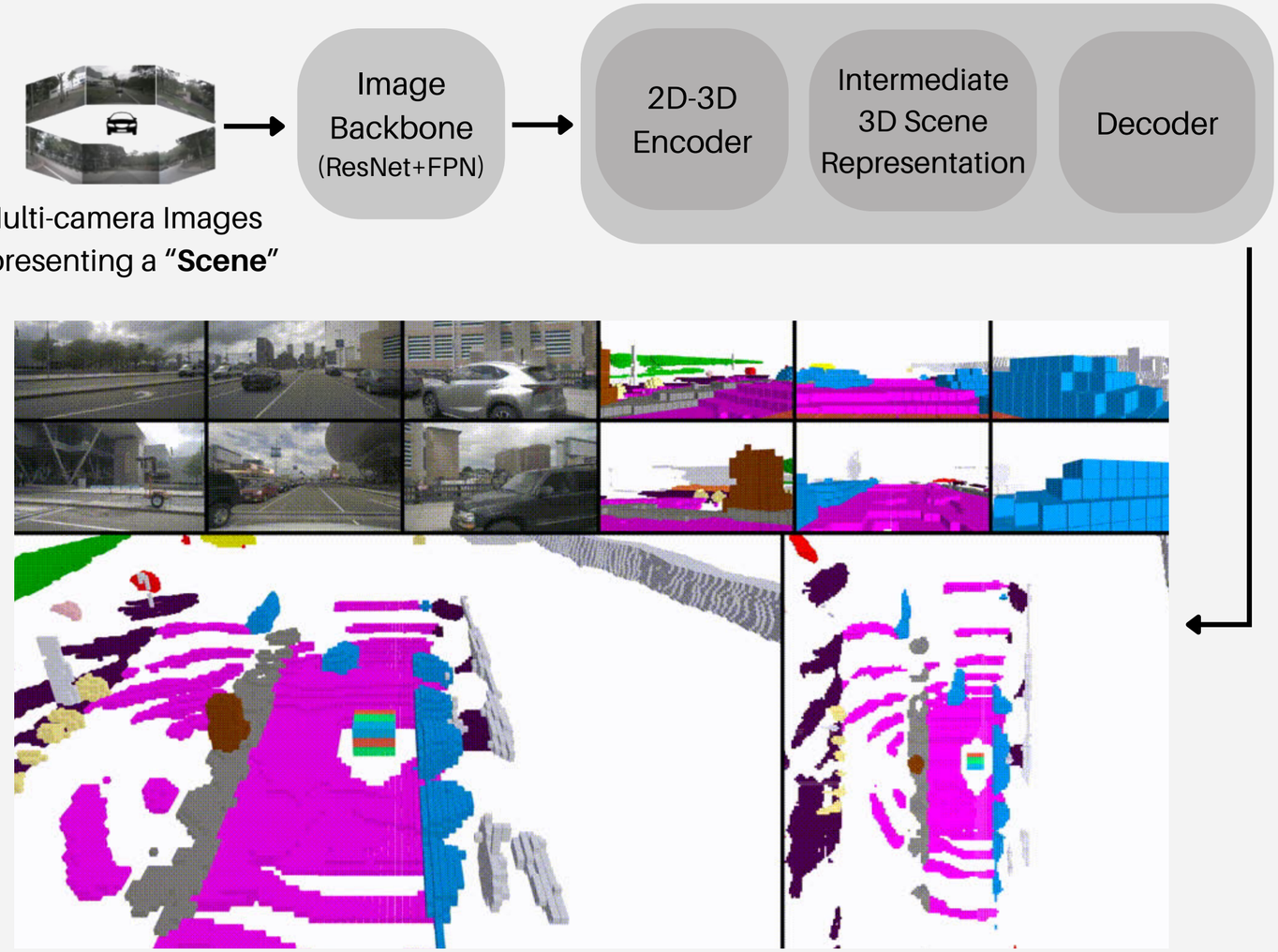
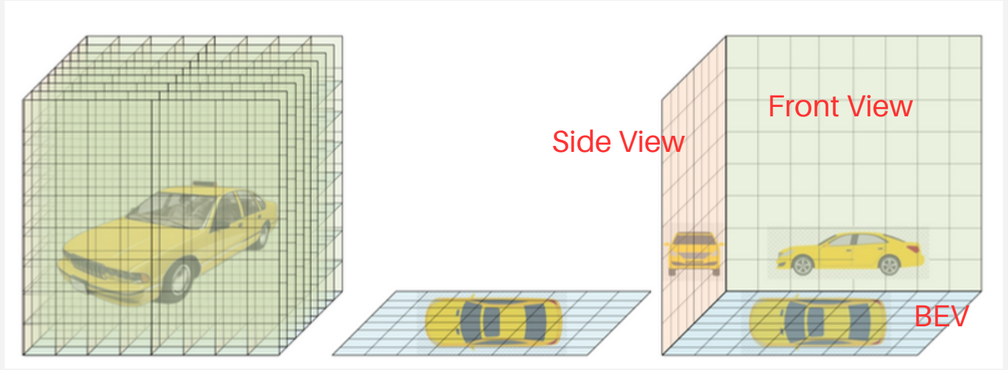


Figure 3: 3D Semantic Occupancy Prediction General Workflow [3]

BEV: Bird's Eye View
TPV: Tri-Perspective View

Related Research

- Our main baseline methods: **BEVFormer, TPVFormer**
  - **BEVFormer**: A ***spatiotemporal transformer*** built for 3D Detection and BEV Map Segmentation
  - **TPVFormer**: Generalizes BEV into *TPV* and uses it for 3D SOP (as well as for SSC + LiDAR Segmentation) --- ***but no temporal information***
- Other: Occ3D, SurroundOcc, Scene as Occupancy



	Voxel	BEV	TPV
Storage/ Computational complexity	$O(HWD)$	$O(HW)$	$O(HW + DH + WD)$

Figure 4: Different Latent Surrounding Scene Representations

Research Questions

- Can we improve TPVFormer's 3D SOP results with **temporal information**? --- introduce **temporal coherence**
  - Make TPVFormer a spatiotemporal transformer --> Inspired by BEVFormer

[4]

BEV: Bird's Eye View
TPV: Tri-Perspective View

Related Research

- Our main baseline methods: **BEVFormer, TPVFormer**
  - **BEVFormer**: A ***spatiotemporal transformer*** built for 3D Detection and BEV Map Segmentation
  - **TPVFormer**: Generalizes BEV into *TPV* and uses it for 3D SOP (as well as for SSC + LiDAR Segmentation) --- ***but no temporal information***
- Other: Occ3D, SurroundOcc, Scene as Occupancy

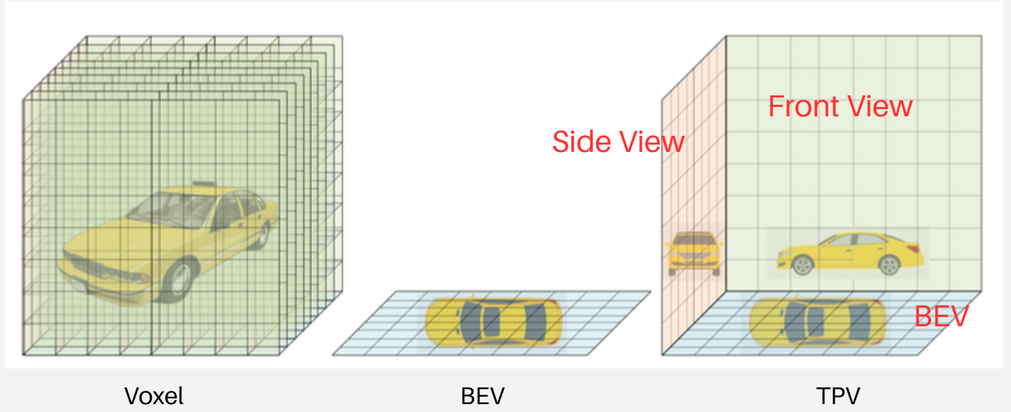


Figure 4: Different Latent Surrounding Scene Representations

Research Questions

- Can we improve TPVFormer's 3D SOP results with **temporal information**? --- introduce **temporal coherence**
  - Make TPVFormer a spatiotemporal transformer --> Inspired by BEVFormer

[4]

BEV: Bird's Eye View
TPV: Tri-Perspective View

Related Research

- Our main baseline methods: **BEVFormer, TPVFormer**
  - **BEVFormer**: A ***spatiotemporal transformer*** built for 3D Detection and BEV Map Segmentation
  - **TPVFormer**: Generalizes BEV into *TPV* and uses it for 3D SOP (as well as for SSC + LiDAR Segmentation) --- ***but no temporal information***
- Other: Occ3D, SurroundOcc, OccFormer

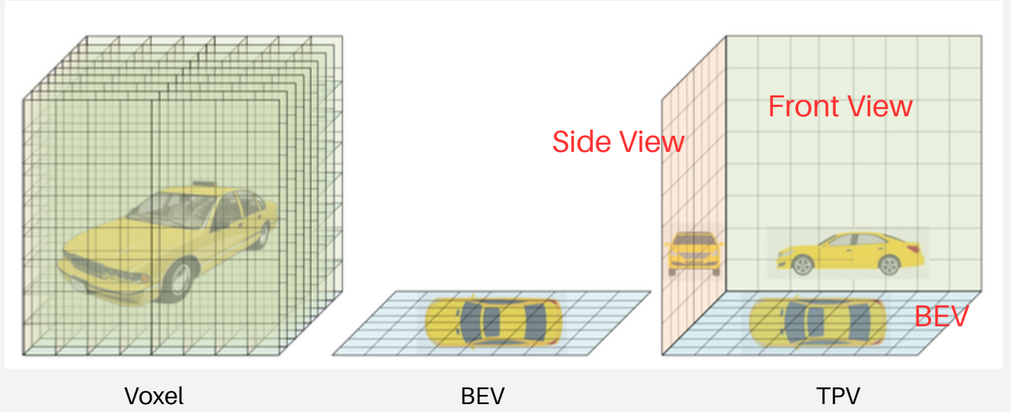


Figure 4: Different Latent Surrounding Scene Representations

Research Questions

- Can we improve TPVFormer's 3D SOP results with **temporal information**? --- introduce **temporal coherence**
  - Make TPVFormer a spatiotemporal transformer --> Inspired by BEVFormer

[4]

BEV: Bird's Eye View
TPV: Tri-Perspective View

Related Research

- Our main baseline methods: **BEVFormer, TPVFormer**
  - **BEVFormer**: A ***spatiotemporal transformer*** built for 3D Detection and BEV Map Segmentation
  - **TPVFormer**: Generalizes BEV into *TPV* and uses it for 3D SOP (as well as for SSC + LiDAR Segmentation) --- ***but no temporal information***
- Other: Occ3D, SurroundOcc, OccFormer

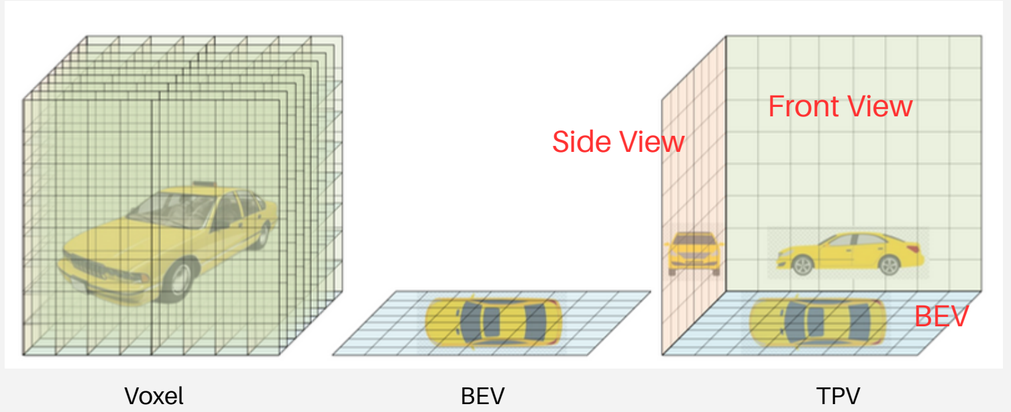


Figure 4: Different Latent Surrounding Scene Representations

Research Questions

- Can we improve TPVFormer's 3D SOP results with **temporal information**? --- introduce **temporal coherence**
  - Make TPVFormer a spatiotemporal transformer --> Inspired by BEVFormer

[4]

nuScenes Dataset

- A public large-scale dataset for autonomous driving
- Contains 1000 driving scenes (~5.5 hours of driving data)
- Other similar datasets: KITTI, Waymo
- Vision-centric SOP uses the 6 camera images surrounding the vehicle for training, and LiDAR points for querying

Figure 5: A visualization of sample LiDAR point cloud in nuScenes dataset [5]

Spatial Reasoning: the ability to think about and manipulate objects in three dimensions

TPVFormer

Sparse supervision ---> Dense inference

- Was the first work to explore the field of SOP.
- Uses the **sparse** LiDAR segmentation annotations provided by the **nuScenes dataset**.

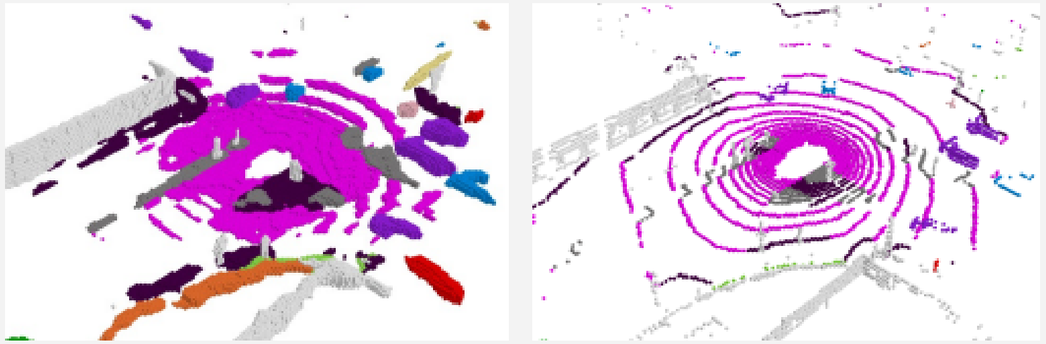


Figure 6: Comparison of the TPVFormer SOP vs Ground Truth

Occ3D, SurroundOCC, Scene as Occupancy

Denser is Better!

- You need high-quality dense annotations for high-quality dense predictions

They are not wrong!
But, we propose a different approach

S2TPVFormer (ours)

Densification of supervision is not the only way for high-quality SOP!

- There's a high chance that the model trying to learn the supervision pipeline (very **obvious in Figure 5**)
- Historical data (temporal) also provide access to learn dense Semantic Occupancy!

[6]

Spatial Reasoning: the ability to think about and manipulate objects in three dimensions

TPVFormer

Sparse supervision ---> Dense inference

- Was the first work to explore the field of SOP.
- Uses the sparse LiDAR segmentation annotations provided by the **nuScenes dataset**.

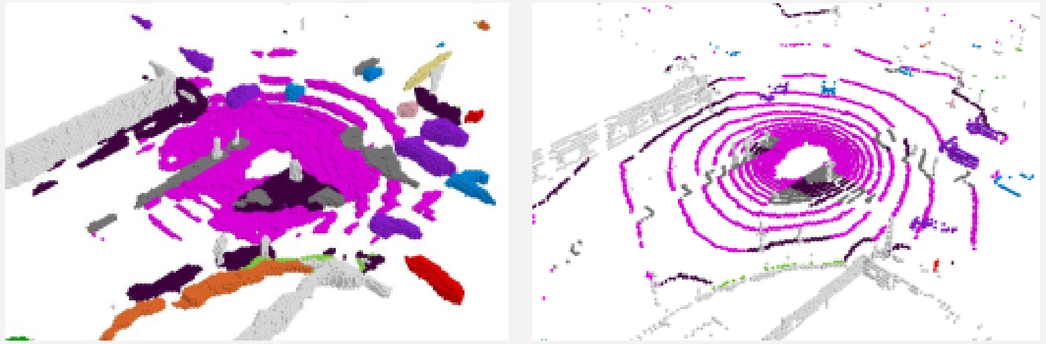


Figure 6: Comparison of the TPVFormer SOP vs Ground Truth

Occ3D, SurroundOCC, Scene as Occupancy

Denser is Better!

- You need high-quality dense annotations for high-quality dense predictions

They are not wrong!
But, we propose a different approach

S2TPVFormer (ours)

Densification of supervision is not the only way for high-quality SOP!

- There's a high chance that the model trying to learn the supervision pipeline (very **obvious in Figure 5**)
- Historical data (temporal) also provide access to learn dense Semantic Occupancy!

[6]

Spatial Reasoning: the ability to think about and manipulate objects in three dimensions

TPVFormer

Sparse supervision ---> Dense inference

- Was the first work to explore the field of SOP.
- Uses the sparse LiDAR segmentation annotations provided by the **nuScenes dataset**.

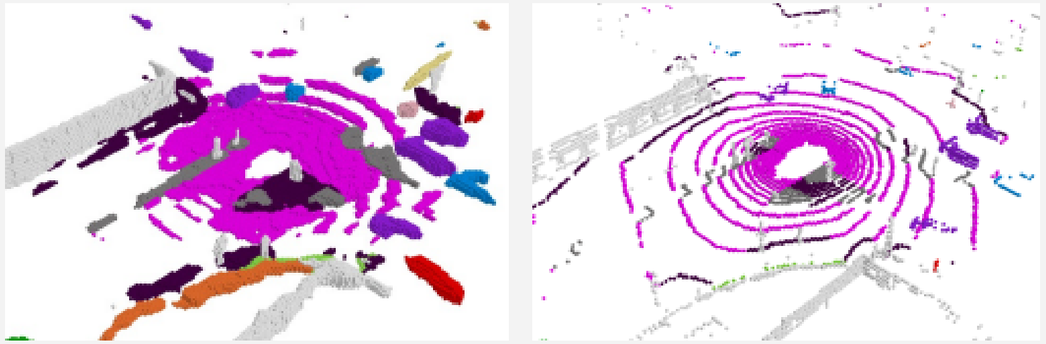


Figure 6: Comparison of the TPVFormer SOP vs Ground Truth

Occ3D, SurroundOCC, Scene as Occupancy

Denser is Better!

- You need high-quality dense annotations for high-quality dense predictions

They are not wrong!
But, we propose a different approach

S2TPVFormer (ours)

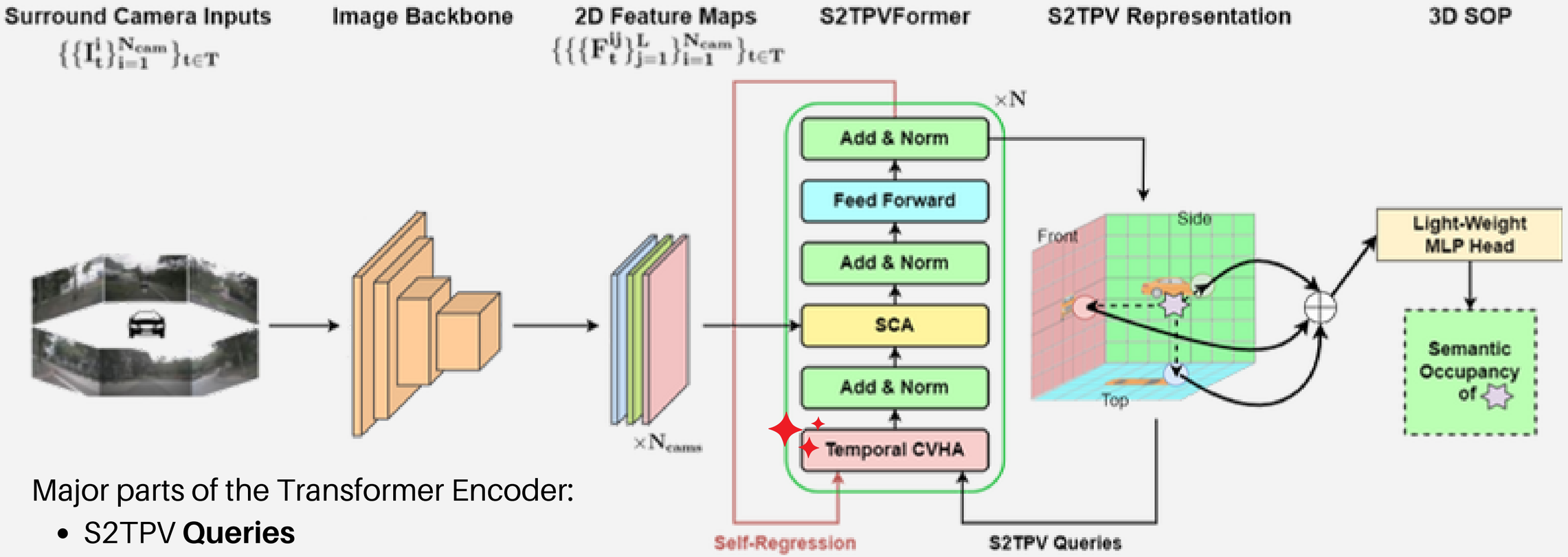
Densification of supervision is not the only way for high-quality SOP!

- There's a high chance that the model trying to learn the supervision pipeline (very **obvious in Figure 5**)
- Historical data (temporal) also provide access to learn dense Semantic Occupancy!

[6]

SCA: Spatial Cross-Attention
CVHA: Cross-View Hybrid Attention

Methodology: S2TPVFormer



Major parts of the Transformer Encoder:

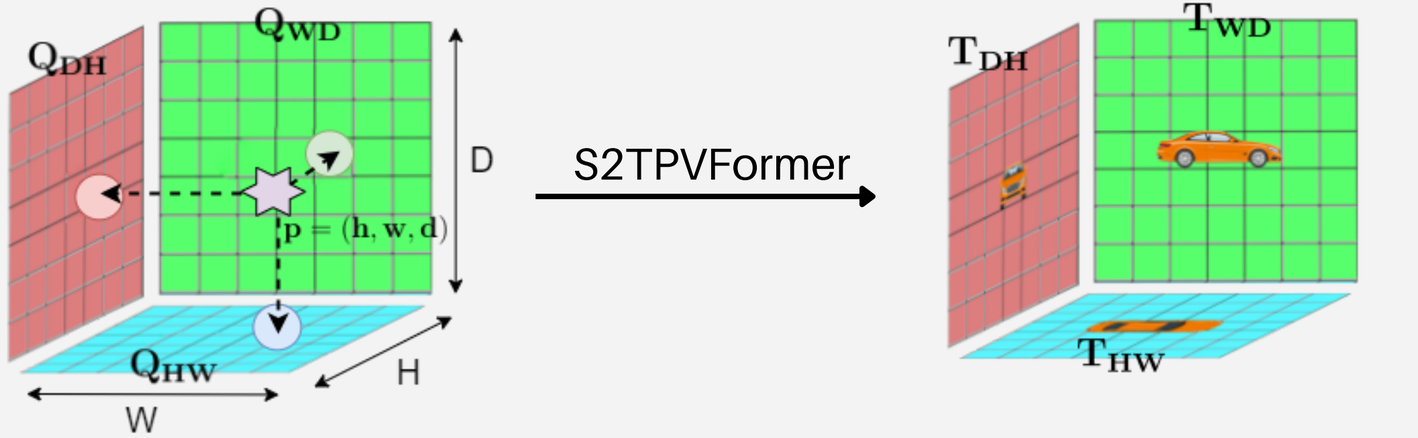
- S2TPV **Queries**
- **SCA** Module
- **Self-Regression** Mechanism
- **Temporal CVHA** Module

Figure 7: S2TPVFormer Architecture and Workflow [7]

S2TPV Queries

Two types of queries:

1. Implicit latent queries
2. Explicit real-world spatial queries



S2TPV Queries

$$\mathbf{Q} = \mathbf{Q_{HW}} \cup \mathbf{Q_{DH}} \cup \mathbf{Q_{WD}}$$

S2TPV Features

$$\mathbf{T} = \mathbf{T_{HW}} \cup \mathbf{T_{DH}} \cup \mathbf{T_{WD}}$$

with spatial resolutions

$$\mathbf{r_{HW}} = 0.512, \mathbf{r_{DH}} = \mathbf{r_{WD}} = 0.5$$

Point queries for a point $\mathbf{p} = (\mathbf{h}, \mathbf{w}, \mathbf{d})$:

$$\mathbf{q_{h,w}} \in \mathbb{R}^{\mathbf{H} \times \mathbf{W} \times \mathbf{C}}$$

$$\mathbf{q_{d,h}} \in \mathbb{R}^{\mathbf{D} \times \mathbf{H} \times \mathbf{C}}$$

$$\mathbf{q_{w,d}} \in \mathbb{R}^{\mathbf{W} \times \mathbf{D} \times \mathbf{C}}$$

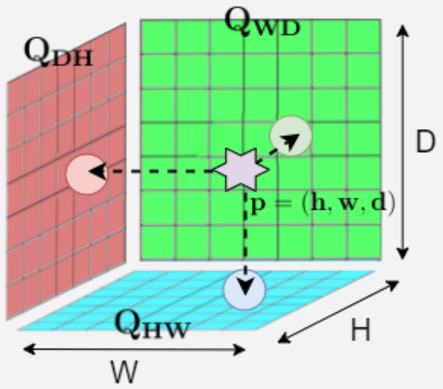
At each encoder layer, the **queries get refined**. In other words, the Encoder layer replaces the queries with its answers, and it refines the answers at each layer.

[8]

Spatial Cross-Attention (SCA)

aka **Spatial Fusion**

Implemented using Deformable Attention

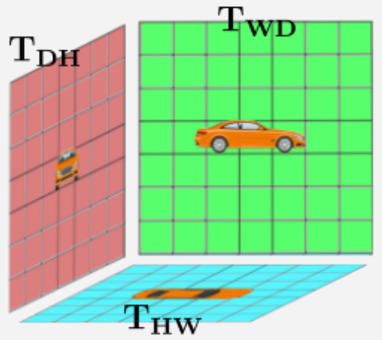


S2TPV Queries

Surround Camera Image Features at timestep **t**

*Fuse the 2D Image features onto the TPV planes

SCA



Intermediate S2TPV Features

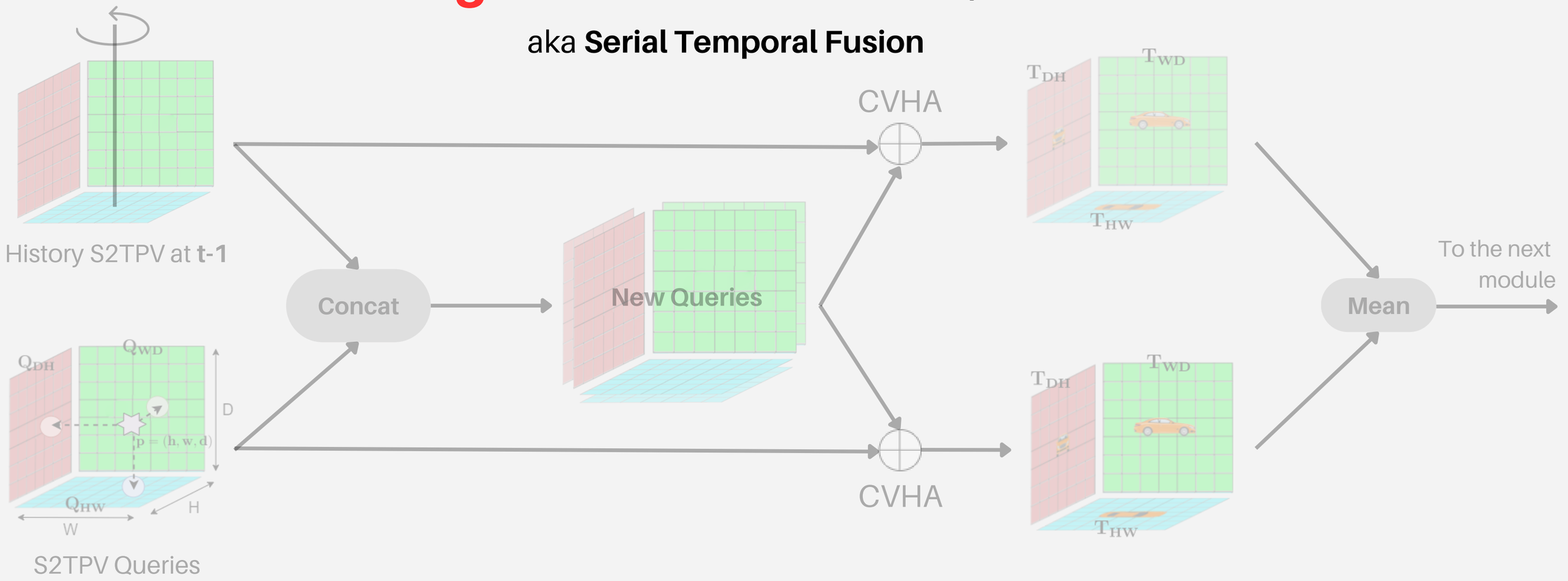
*Camera intrinsics & extrinsics provide geometry prior for the 2D to 3D fusion to happen

Temporal Cross-View Hybrid Attention & Self-Regression

aka **Serial Temporal Fusion**

Our main contribution

Implemented using Deformable Attention



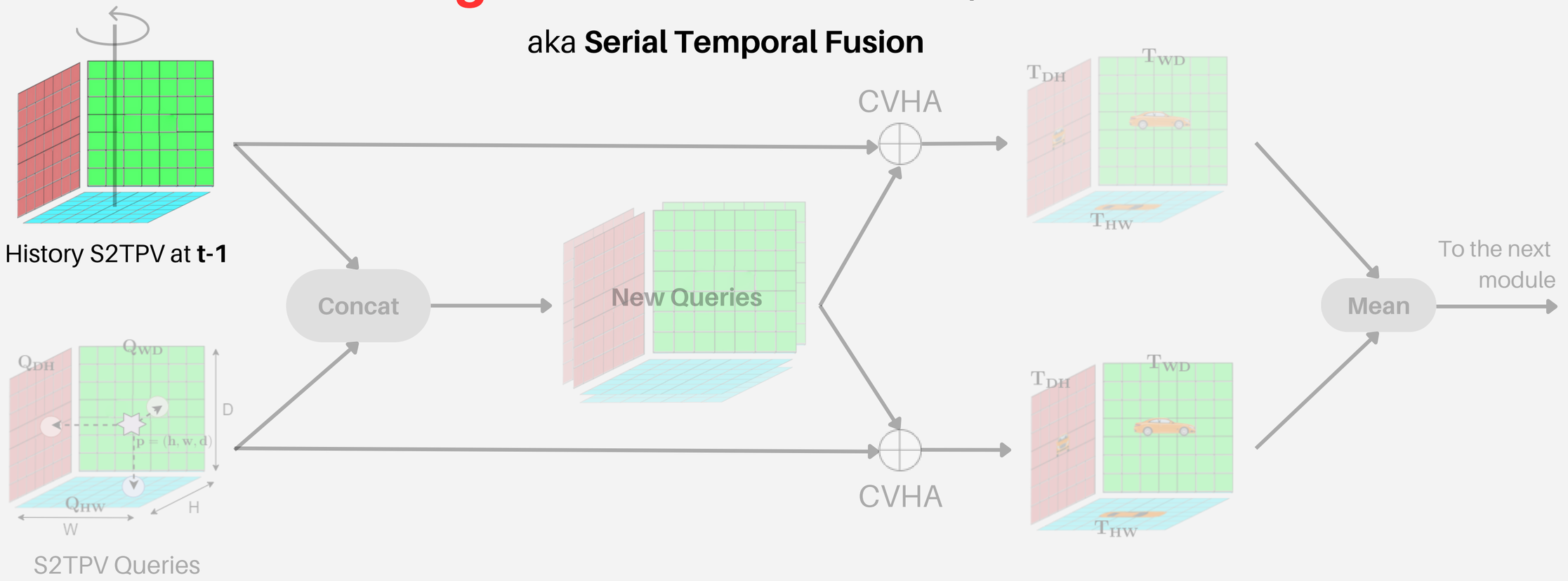
Previous history S2TPV features (at **t-1**) are obtained by **re-running** the transformer encoder recursively, with the new queries as input, which realizes a **self-regression** mechanism

Temporal Cross-View Hybrid Attention & Self-Regression

aka **Serial Temporal Fusion**

Our main contribution

Implemented using Deformable Attention



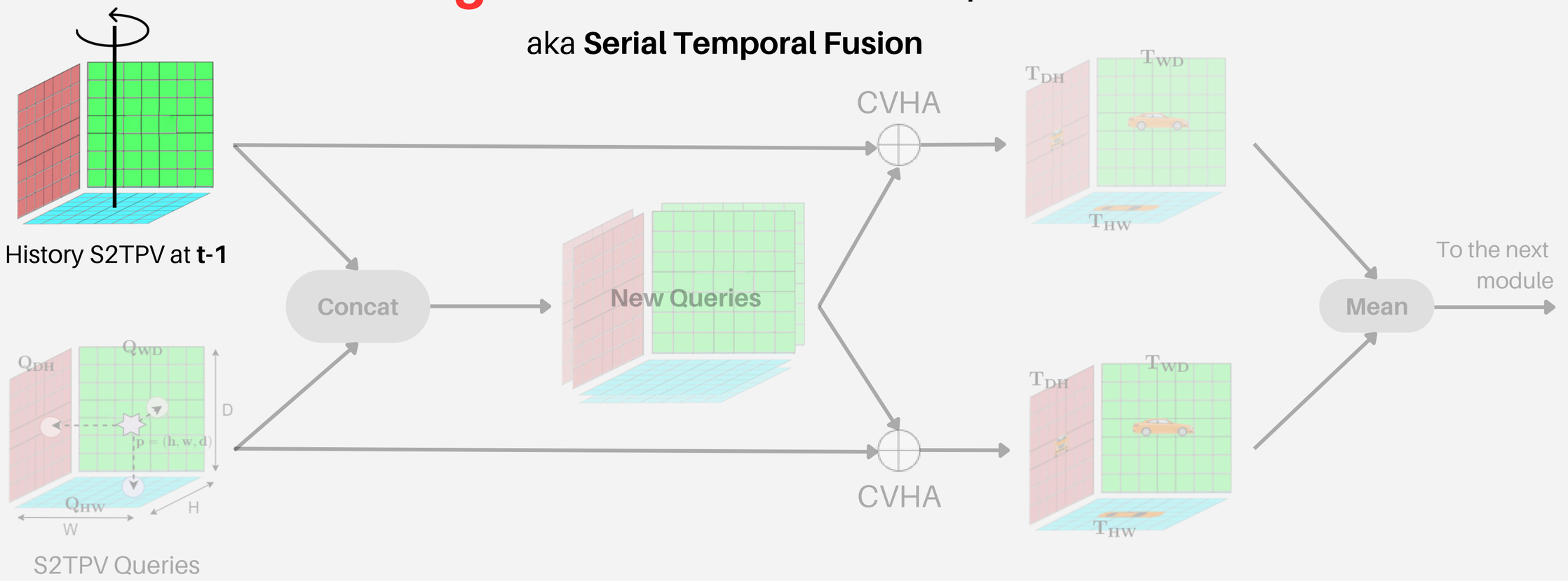
Previous history S2TPV features (at **t-1**) are obtained by **re-running** the transformer encoder recursively, with the new queries as input, which realizes a **self-regression** mechanism

Temporal Cross-View Hybrid Attention & Self-Regression

Our main contribution

Implemented using Deformable Attention

aka **Serial Temporal Fusion**



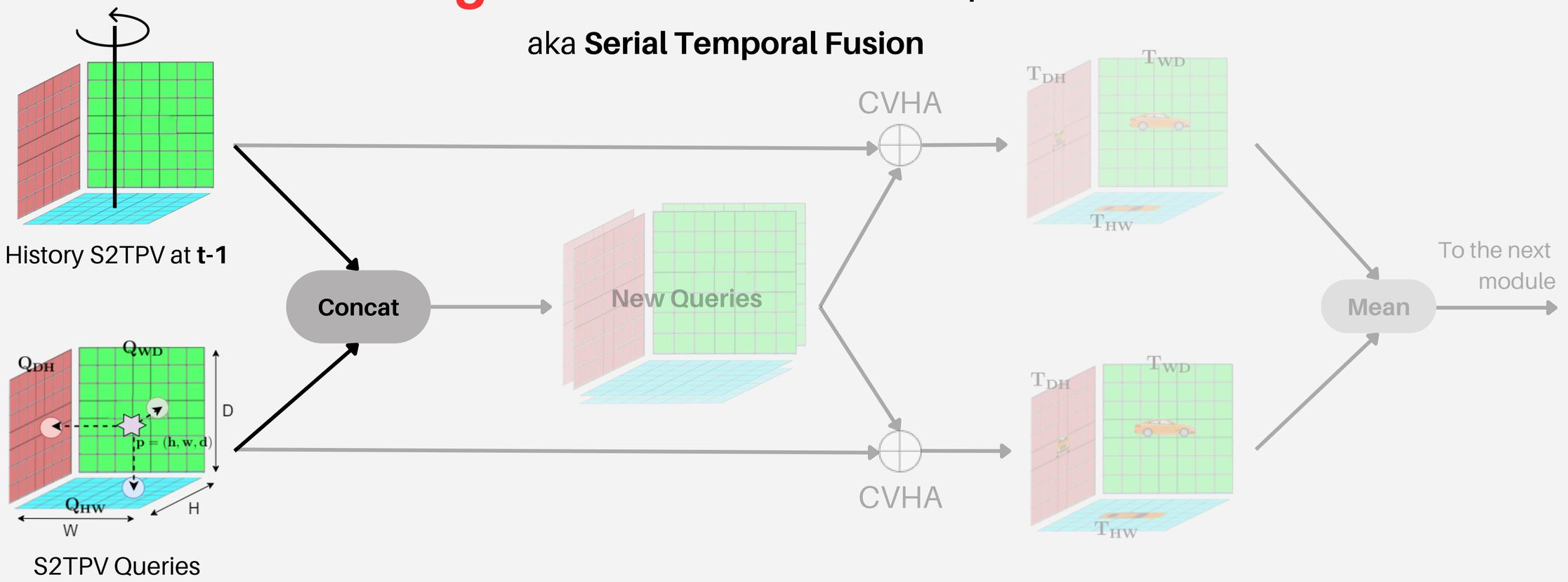
Previous history S2TPV features (at **t-1**) are obtained by **re-running** the transformer encoder recursively, with the new queries as input, which realizes a **self-regression** mechanism

Temporal Cross-View Hybrid Attention & Self-Regression

Our main contribution

Implemented using Deformable Attention

aka **Serial Temporal Fusion**



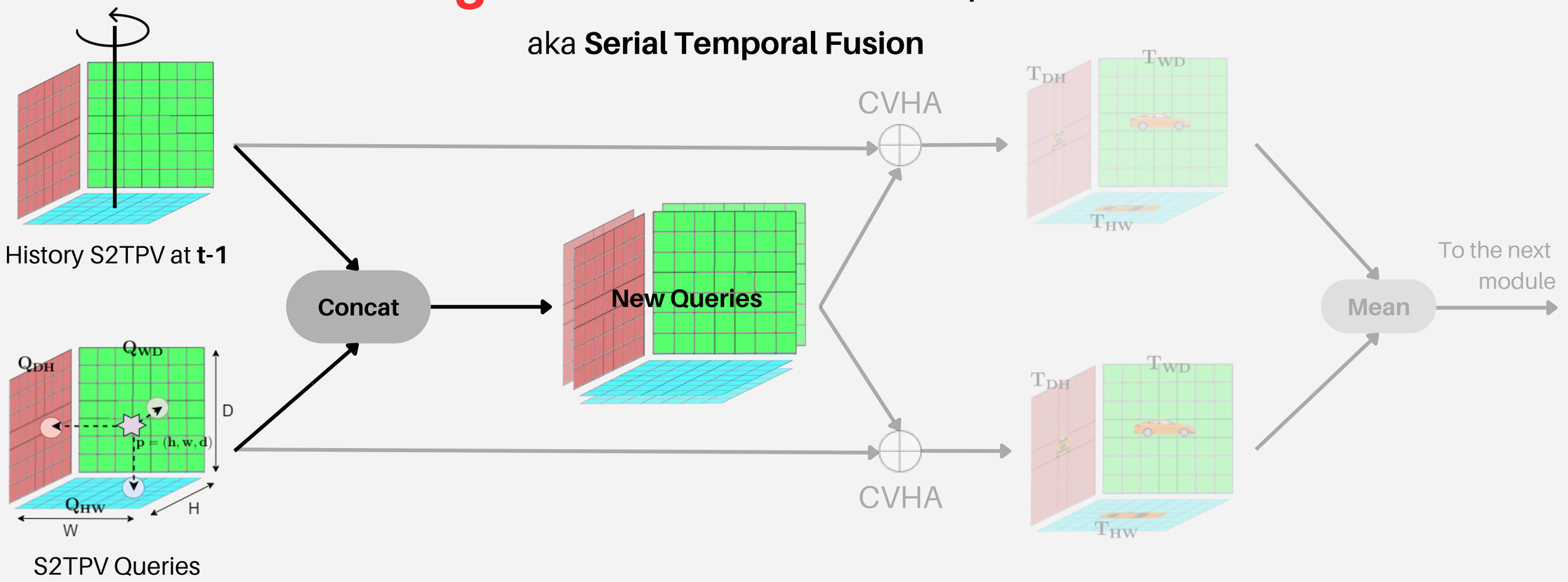
Previous history S2TPV features (at **t-1**) are obtained by **re-running** the transformer encoder recursively, with the new queries as input, which realizes a **self-regression** mechanism

Temporal Cross-View Hybrid Attention & Self-Regression

Our main contribution

Implemented using Deformable Attention

aka **Serial Temporal Fusion**



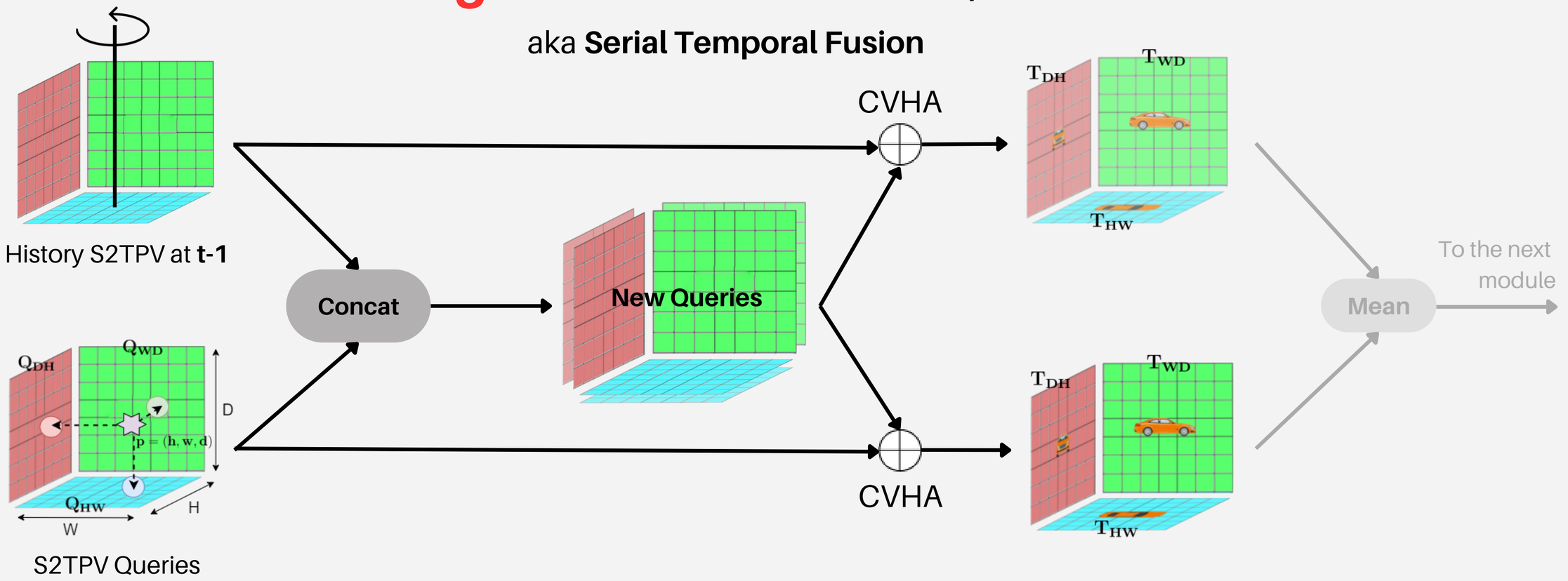
Previous history S2TPV features (at **t-1**) are obtained by **re-running** the transformer encoder recursively, with the new queries as input, which realizes a **self-regression** mechanism

Temporal Cross-View Hybrid Attention & Self-Regression

Our main contribution

Implemented using Deformable Attention

aka **Serial Temporal Fusion**



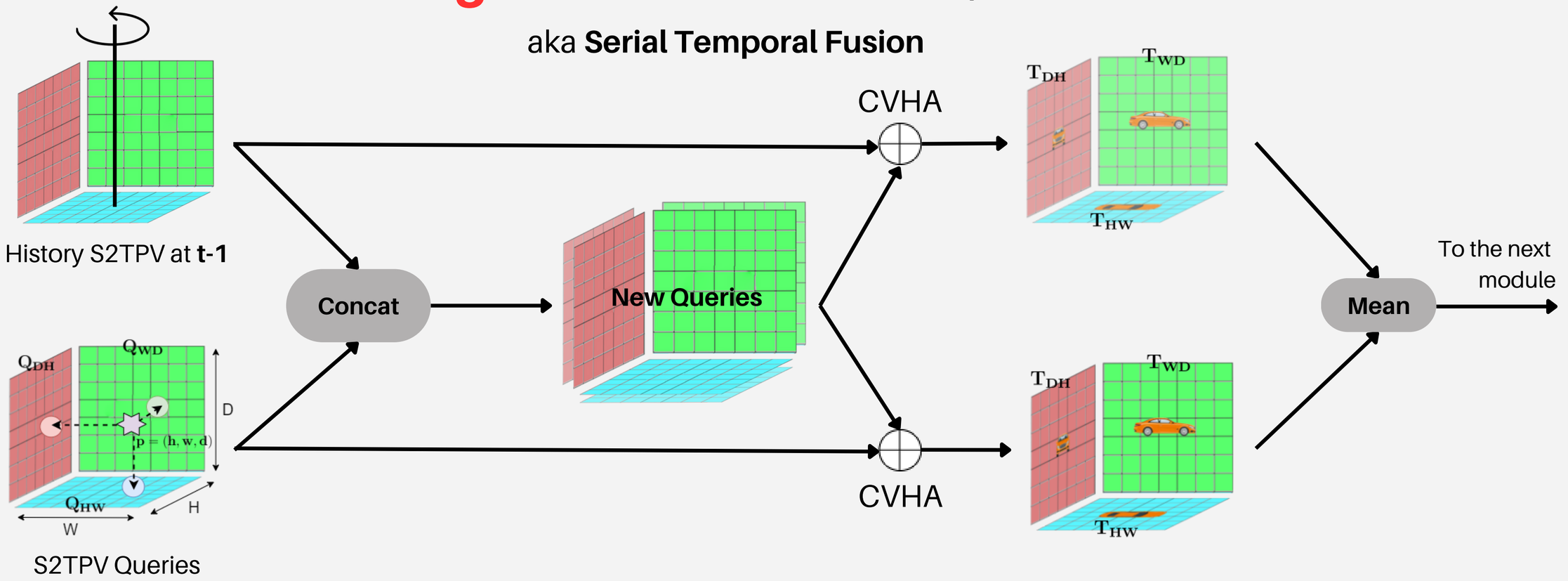
Previous history S2TPV features (at **t-1**) are obtained by **re-running** the transformer encoder recursively, with the new queries as input, which realizes a **self-regression** mechanism

Temporal Cross-View Hybrid Attention & Self-Regression

Our main contribution

Implemented using Deformable Attention

aka **Serial Temporal Fusion**



Previous history S2TPV features (at **t-1**) are obtained by **re-running** the transformer encoder recursively, with the new queries as input, which realizes a **self-regression** mechanism

Experiments - What? & Why?

1. **Base 3D SOP experiments** with small configurations -- to get results quickly

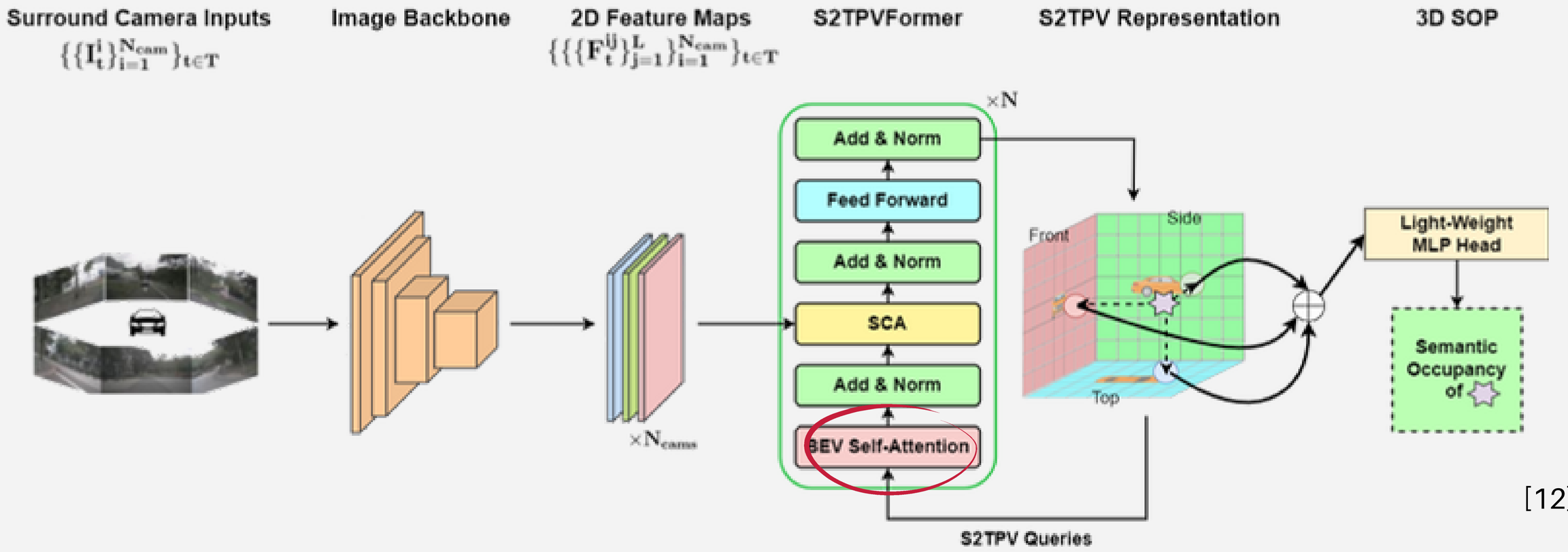
Model	CVHA	Temp	Embed_Dim	TPV_Res	Backbone	Enc_Layers
TPVFormer04	CVHA_False	Temp_False	128	100x100x8	ResNet50	3xSHCAB
S2TPVFormer01	CVHA_False	Temp_False	128	100x100x8	ResNet50	3xSHCAB
S2TPVFormer02	CVHA_True	Temp_False	128	100x100x8	ResNet50	3xSHCAB
S2TPVFormer03	CVHA_False	Temp_BEV	128	100x100x8	ResNet50	3xSHCAB
S2TPVFormer06	CVHA_True	Temp_BEV	128	100x100x8	ResNet50	3xSHCAB
S2TPVFormer04	CVHA_True	Temp_TPV	128	100x100x8	ResNet50	3xSHCAB

2. 3D SOP experiments for **ablations studies**

Model	CVHA	Temp	Embed_Dim	TPV_Res	Backbone	Enc_Layers	Time_Range
exp_abl_1	CVHA_False	Temp_BEV	256	100x100x8	ResNet50	3xTHCAB	3s_4t
exp_abl_2	CVHA_False	Temp_BEV	128	100x100x8	ResNet101	3xTHCAB	3s_4t
exp_abl_3	CVHA_False	Temp_BEV	128	200x200x16	ResNet50	3xTHCAB	3s_4t
exp_abl_4	CVHA_True	Temp_BEV	128	100x100x8	ResNet50	(3xHCAB+2xHAB)	3s_4t
exp_abl_5	CVHA_True	Temp_TPV	256	100x100x8	ResNet50	3xTHCAB	3s_4t
exp_abl_6	CVHA_False	Temp_BEV	128	100x100x8	ResNet50	3xTHCAB	6s_8t

Experiments - Results

1. SOP experiments with small configurations -- to get results quickly
    a. **S2TPVFormer01** = (**CVHA_False**, **Temp_False**, 128, 100x100x8, ResNet50, 3xSHCAB)



[12]

Experiments - Results

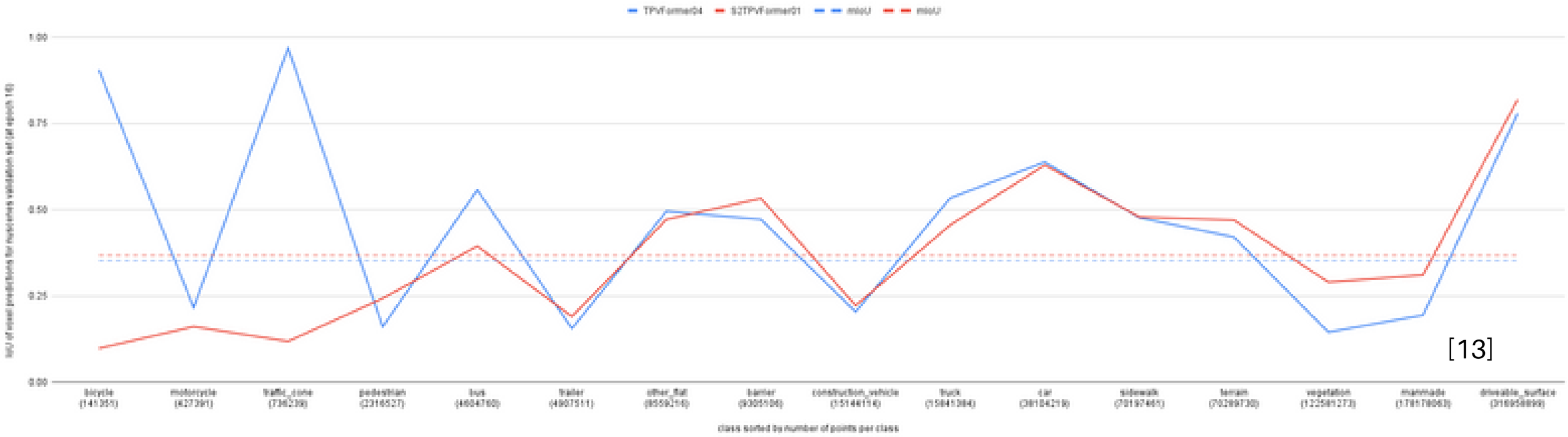
IoU (Intersection over Union) is a measure of overlap between predicted voxels and ground-truth voxels

1. SOP experiments with small configurations -- to get results quickly
    a. **S2TPVFormer01** = (**CVHA_False**, **Temp_False**, 128, 100x100x8, ResNet50, 3xSHCAB)

- Conclusion:
    - Our re-implementation (S2TPVformer) has the same functionality of TPVFormer04

mIoU scores at **3 epochs**:

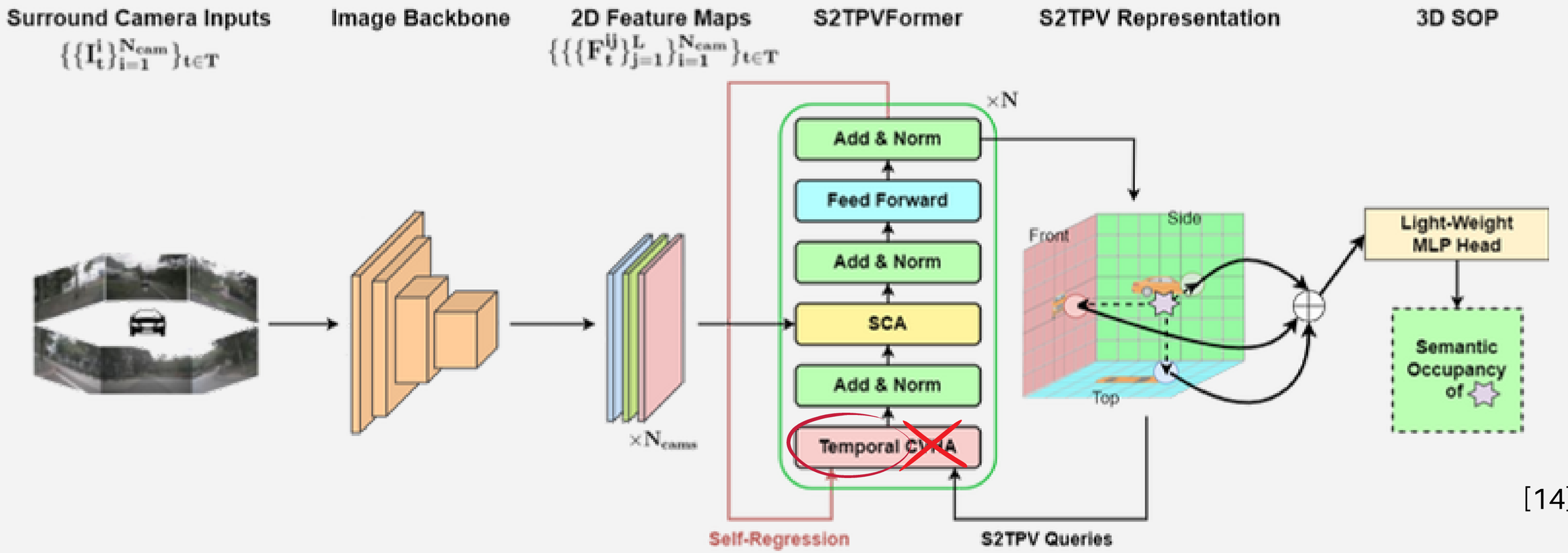
- TPVFormer04 = 35.23 %
- S2TPVFormer01 = **36.79 %**



[13]

Experiments - Results

1. SOP experiments with small configurations -- to get results quickly
    c. **S2TPVFormer03** = (**CVHA_False**, **Temp_BEV**, 128, 100x100x8, ResNet50, 3xTHCAB)



[14]

Experiments - Results

IoU (Intersection over Union) is a measure of overlap between predicted voxels and ground-truth voxels

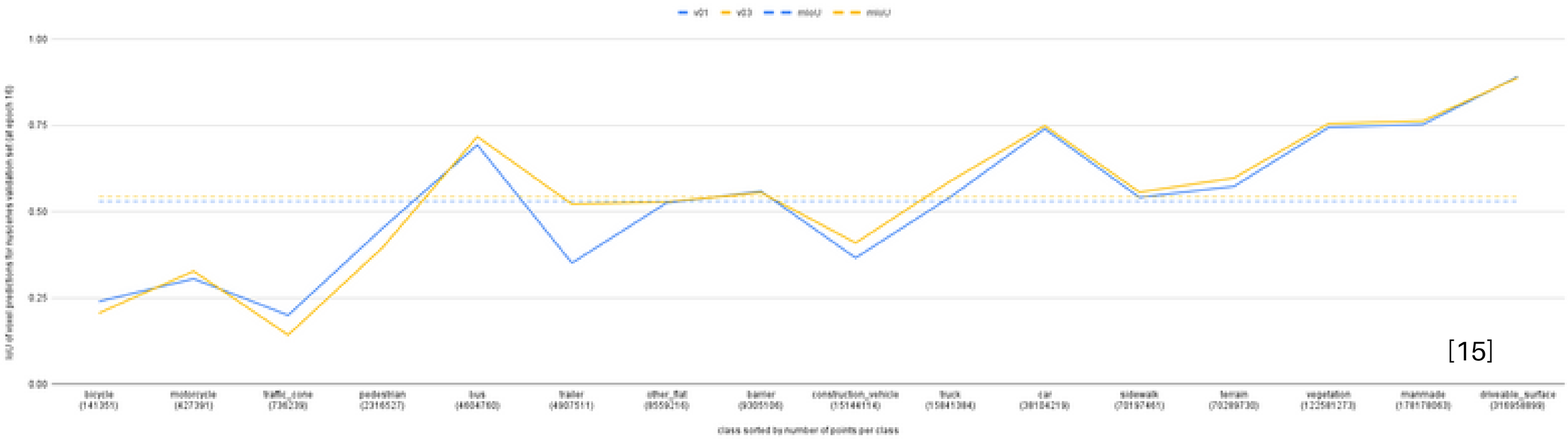
1. SOP experiments with small configurations -- to get results quickly
    c. **S2TPVFormer03** = (**CVHA_False**, **Temp_BEV**, 128, 100x100x8, ResNet50, 3xTHCAB)

- Conclusion:
    - Temporal information helps the model to learn better

mIoU scores at **7 epochs**:

from literature ⇨
- S2TPVFormer01 = 52.99 %

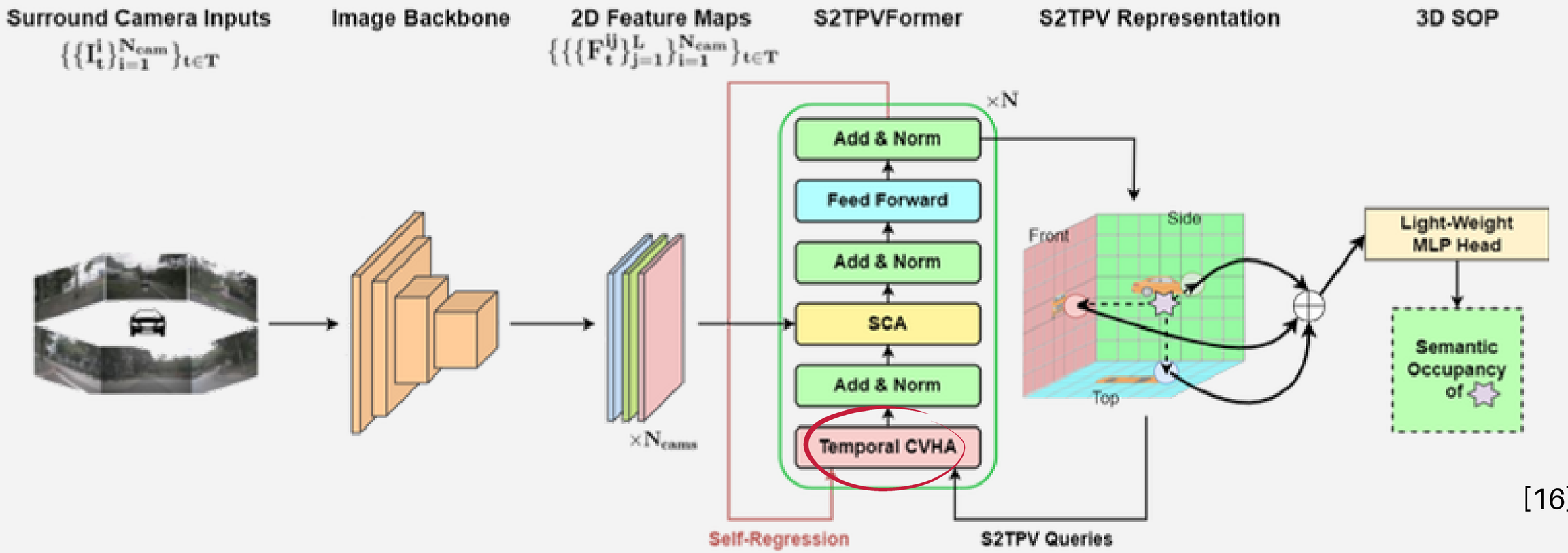
ours ⇨
- S2TPVFormer03 = **54.41 %**
    - diff = **+1.42 %**



[15]

Experiments - Results

1. SOP experiments with small configurations -- to get results quickly
    d. **S2TPVFormer04 =** (**CVHA_True**, **Temp_TPV**, 128, 100x100x8, ResNet50, 3xHCAB)



[16]

Experiments - Results

IoU (Intersection over Union) is a measure of overlap between predicted voxels and ground-truth voxels

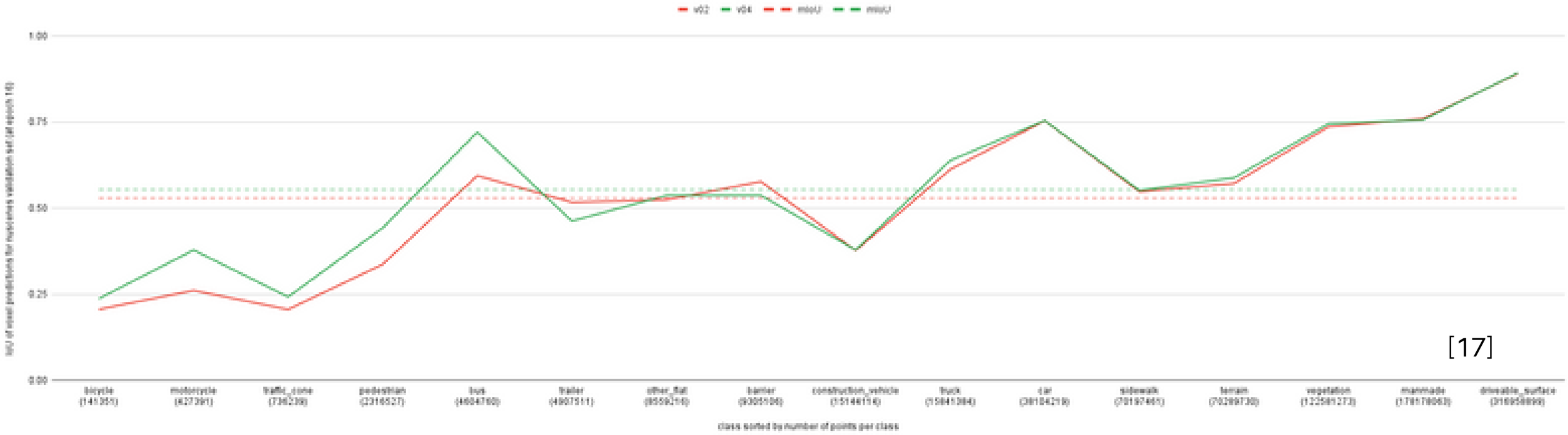
1. SOP experiments with small configurations -- to get results quickly
    d. **S2TPVFormer04 =** (**CVHA_True**, **Temp_TPV**, 128, 100x100x8, ResNet50, 3xHCAB)

- Conclusion:
    - Temporal attention + CVHA works better together.

mIoU scores at **7 epochs**:

from literature ⇨
- S2TPVFormer02 = 52.95 %

ours ⇨
- S2TPVFormer04 = **55.40 %**
    - diff = **+2.45 %**



[17]

Experiments - Results

IoU (Intersection over Union) is a measure of overlap between predicted voxels and ground-truth voxels

1. SOP experiments with small configurations -- to get results quickly

Overall results

mIoU scores at **7 epochs**:

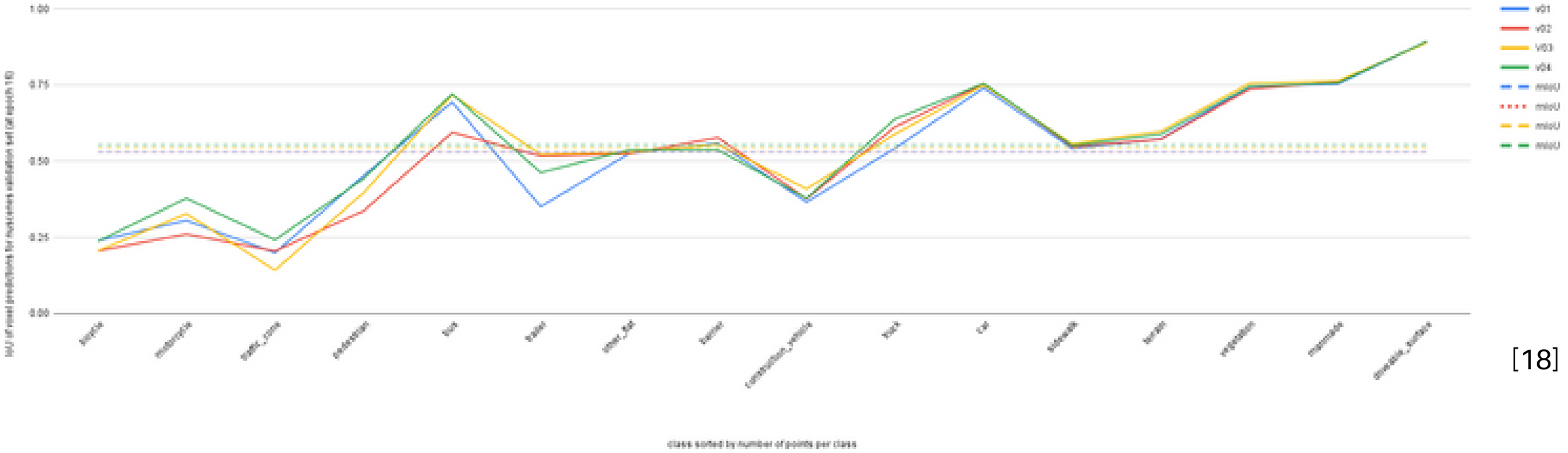
from literature ⇨ • S2TPVFormer01 = 52.99 %

from literature ⇨ • S2TPVFormer02 = 52.95 % } **-0.04 %**

ours ⇨ • S2TPVFormer03 = **54.41 %** } **+1.46 %**

ours ⇨ • S2TPVFormer04 = **55.40 %** } **+0.99 %**

- Conclusion:
  - Contribution to performance improvement:
    - **(TA + CVHA) > TA > CVHA**



[18]

Experiments - What? & Why?

1. **SOP experiments** with small configurations -- to get results quickly

Model	CVHA	Temp	Embed_Dim	TPV_Res	Backbone	Enc_Layers
TPVFormer04	CVHA_False	Temp_False	128	100x100x8	ResNet50	3xSHCAB
S2TPVFormer01	CVHA_False	Temp_False	128	100x100x8	ResNet50	3xSHCAB
S2TPVFormer02	CVHA_True	Temp_False	128	100x100x8	ResNet50	3xSHCAB
S2TPVFormer03	CVHA_False	Temp_BEV	128	100x100x8	ResNet50	3xSHCAB
S2TPVFormer06	CVHA_True	Temp_BEV	128	100x100x8	ResNet50	3xSHCAB
S2TPVFormer04	CVHA_True	Temp_TPV	128	100x100x8	ResNet50	3xSHCAB

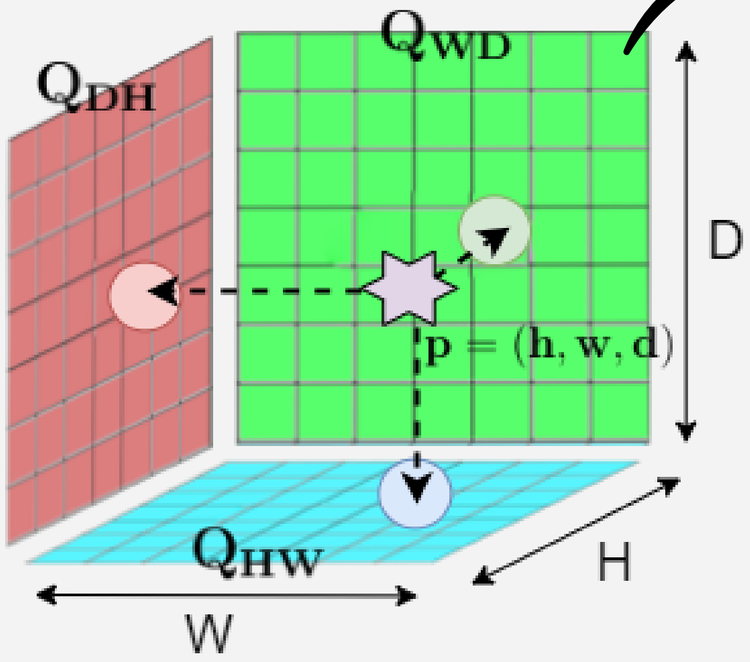
2. 3D SOP experiments for **ablations studies**

Model	CVHA	Temp	Embed_Dim	TPV_Res	Backbone	Enc_Layers	Time_Range
exp_abl_1	CVHA_False	Temp_BEV	256	100x100x8	ResNet50	3xTHCAB	3s_4t
exp_abl_2	CVHA_False	Temp_BEV	128	100x100x8	ResNet101	3xTHCAB	3s_4t
exp_abl_3	CVHA_False	Temp_BEV	128	200x200x16	ResNet50	3xTHCAB	3s_4t
exp_abl_4	CVHA_True	Temp_BEV	128	100x100x8	ResNet50	(3xHCAB+2xHAB)	3s_4t
exp_abl_5	CVHA_True	Temp_TPV	256	100x100x8	ResNet50	3xTHCAB	3s_4t
exp_abl_6	CVHA_False	Temp_BEV	128	100x100x8	ResNet50	3xTHCAB	6s_8t

Experiments - Results

2. SOP experiments for ablations studies
    a. **exp_abl_1** = (CVHA_False, Temp_BEV, **256**, 100x100x8, ResNet50, 3xTHCAB) -- **from S2TPVFormer03**
    embed_dim: **128** --> **256**

dimensionality of Q(h,w,d)



[20]

IoU (Intersection over Union) is a measure of overlap between predicted voxels and ground-truth voxels

Experiments - Results

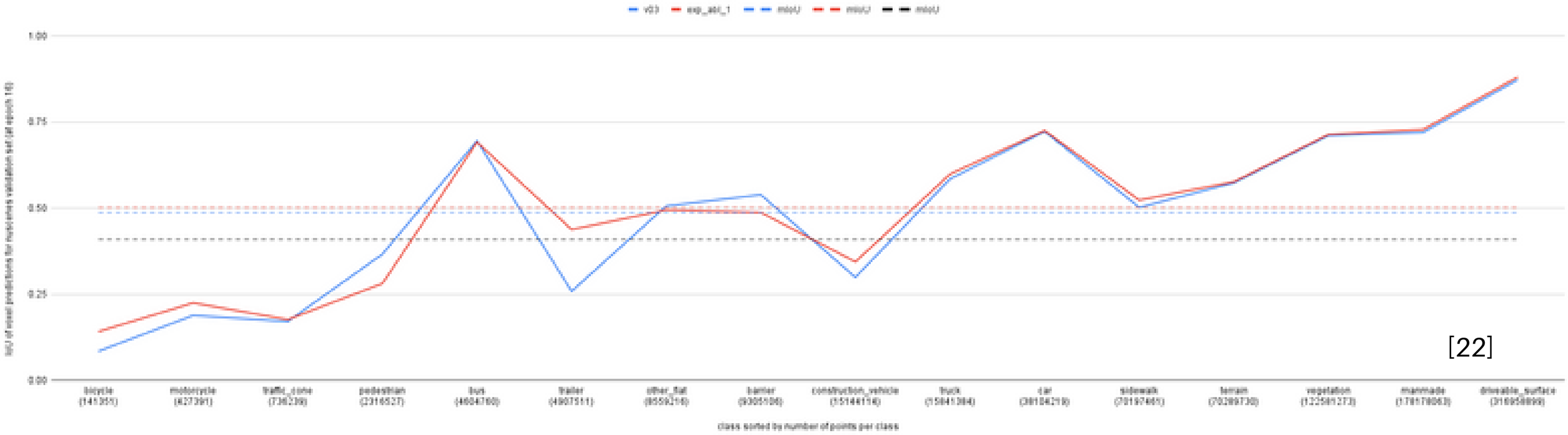
2. SOP experiments for ablations studies
   a. **exp_abl_1** = (CVHA_False, Temp_BEV, **256**, 100x100x8, ResNet50, 3xTHCAB) -- **from S2TPVFormer03**
   embed_dim: **128** --> **256**
- Conclusion:
  - Experiment shows better results for a;
    - **Higher embed dimensionality**

mIoU scores at **1 epochs**:

from literature ⇨
- S2TPVFormer01 = 40.92 %

ours ⇨
- S2TPVFormer03 = 48.73 %

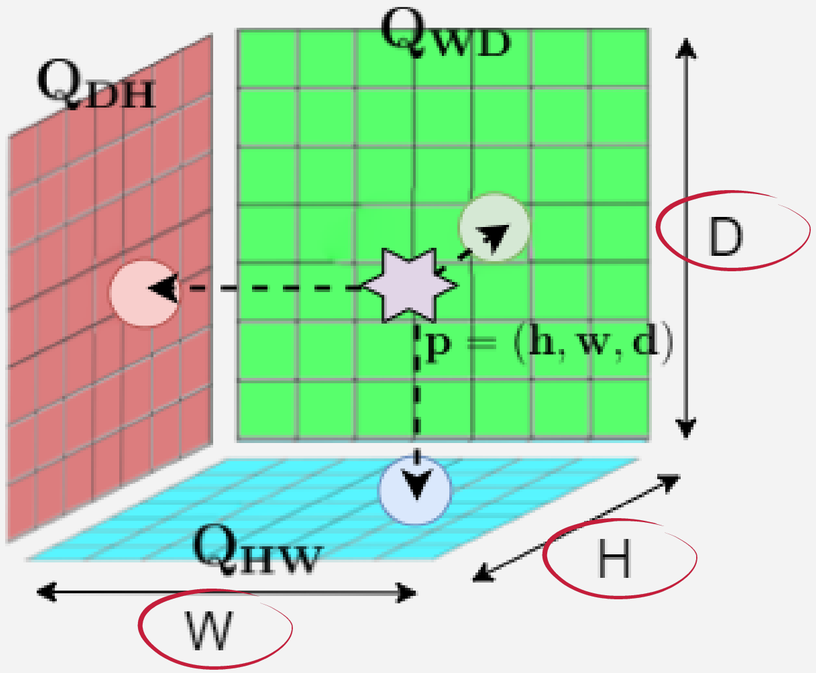
ours ⇨
- exp_abl_1 = **50.18 %**
  - diff = **+1.45 %**



[22]

Experiments - Results

2. SOP experiments for ablations studies
   c. **exp_abl_3** = (CVHA_False, Temp_BEV, 128, **200x200x16**, ResNet50, 3xTHCAB) -- **from S2TPVFormer03**
      TPV_Res: **100x100x8** --> **200x200x16**



Values of H, W and D

[23]

IoU (Intersection over Union) is a measure of overlap between predicted voxels and ground-truth voxels

Experiments - Results

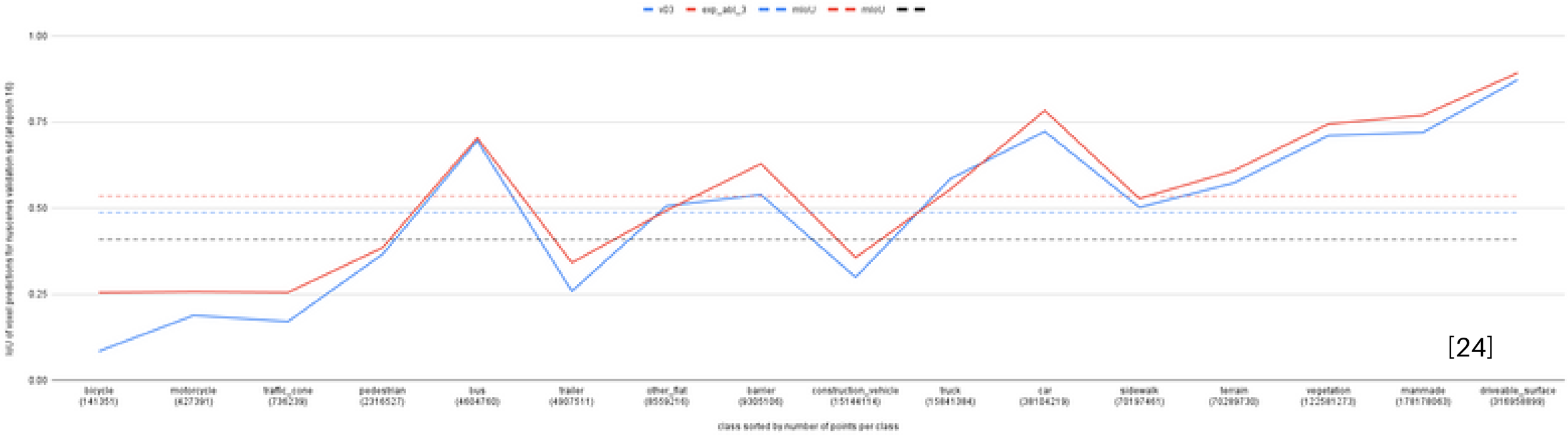
2. SOP experiments for ablations studies
    c. **exp_abl_3** = (CVHA_False, Temp_BEV, 128, **200x200x16**, ResNet50, 3xTHCAB) -- **from S2TPVFormer03**
    TPV_Res: **100x100x8** --> **200x200x16**
- Conclusion:
    - Experiment shows better results for a;
        - **Higher TPV resolution**

mIoU scores at **1 epochs**:

from literature ⇨
- S2TPVFormer01 = 40.92 %

ours ⇨
- S2TPVFormer03 = 48.73 %

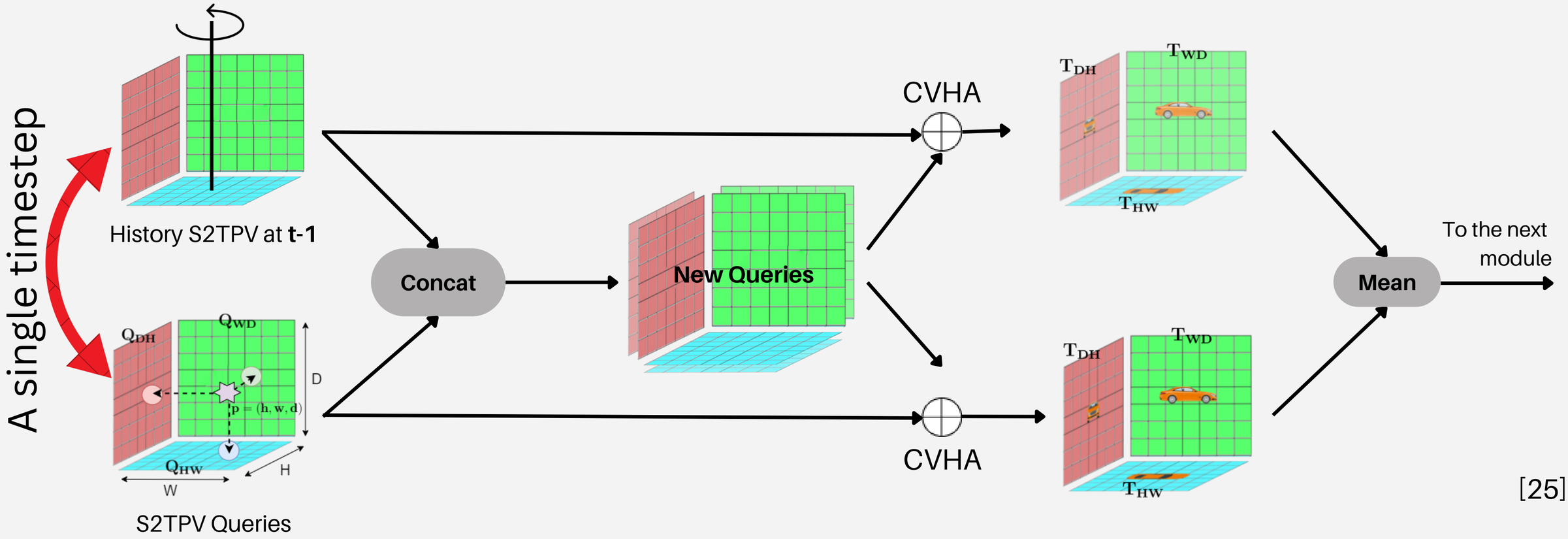
ours ⇨
- exp_abl_3 = **53.49 %**
    - diff = **+4.76 %**



[24]

Experiments - Results

2. SOP experiments for ablations studies
   f. **exp_abl_6** = (CVHA_False, Temp_BEV, 128, 100x100x8, ResNet50, 3xTHCAB, **6s_8t**) -- **from S2TPVFormer03**
   No. past time frames: **3s_4t** --> **6s_8t**



[25]

IoU (Intersection over Union) is a measure of overlap between predicted voxels and ground-truth voxels

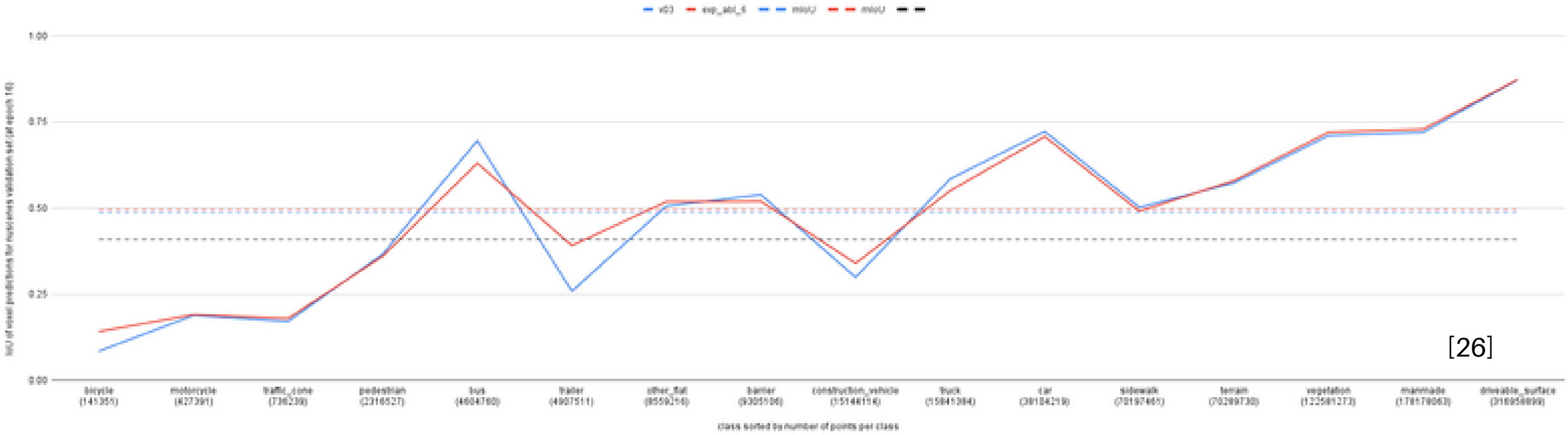
Experiments - Results

2. SOP experiments for ablations studies
   f. **exp_abl_6** = (CVHA_False, Temp_BEV, 128, 100x100x8, ResNet50, 3xTHCAB, **6s_8t**) -- **from S2TPVFormer03**
   No. past time frames: **3s_4t** --> **6s_8t**

- Conclusion:
  - Experiment shows better results for a;
    - **Longer range temporal fusion**.
  - Note: Need to do more experiments to understand the actual trend.

mIoU scores at **1 epochs**:

- *from literature* ⇨ S2TPVFormer01 = 40.92 %
- *ours* ⇨ S2TPVFormer03 = 48.73 %
- *ours* ⇨ exp_abl_3 = **49.55 %**
  - diff = **+0.82 %**



[26]

Experiments - Results

IoU (Intersection over Union) is a measure of overlap between predicted voxels and ground-truth voxels

2. SOP experiments for ablations studies

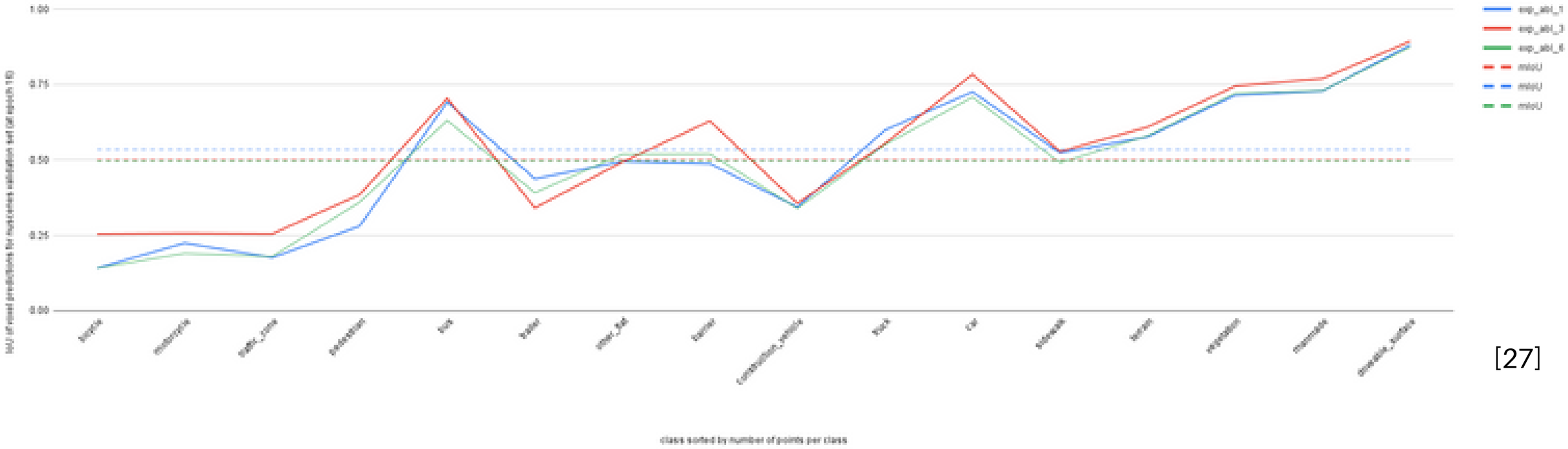
Overall results

- Conclusion:
  - Contribution to performance improvement: Increasing;
    - **TPV resolution > Embed. dim. > Range of TA**

(exp_abl_3) (exp_abl_1) (exp_abl_6)

mIoU scores at **1 epochs**:

- exp_abl_6 = 49.55 %
- exp_abl_1 = 50.18 % } **+0.63 %**
- exp_abl_3 = 53.49 % } **+3.31 %**



[27]

Deliverables

1. Beating the per class IoU scores of TPVFormer04 with the improved S2TPVFormer03 architecture

2. Beating other state of the art 3D SOP models with the best hyper-parameter tuned S2TPVFormer

- **3D Occupancy Prediction leaderboard** of the **CVPR 2023 Autonomous Driving Challenge**
- **3D Occupancy Prediction challenge** on **EvalAI**
  - EvalAI is an open source platform for evaluating and comparing machine learning (ML) and artificial intelligence (AI) algorithms at scale.

Use the ablation studies to find out the set of hyper parameters that gives the best performance

Train the model until we see an increase in validation metrics

Submit the model with the best hyper-parameters and checkpoint to open challenges/leaderboards

[28]

Rank	Participant team	mIoU (↑)	others (↑)	barrier (↑)	bicycle (↑)	bus (↑)	car (↑)	construction_vehicle (↑)	motorcycle (↑)	pedestriar (↑)
1	Noah Ark's Lab (MetaGOD)	54.76	25.65	58.85	44.42	57.54	63.27	38.63	51.50	52.75
2	NVOCC (FB-OCC)	54.19	28.95	57.98	46.40	52.36	63.07	35.68	48.81	42.98
3	42dot (MiLO)	52.45	27.80	56.28	42.62	50.27	61.01	35.41	47.97	38.90
4	UniOcc (final)	51.27	26.94	56.17	39.55	49.40	60.42	35.51	44.77	42.96
5	occ-heiheihei	49.36	28.43	54.49	39.04	45.45	59.15	32.05	43.46	36.33
6	occ_transformer	49.23	26.91	53.57	39.53	47.56	59.54	32.59	44.34	37.36
7	CakeCake (Noah CV Lab - POP)	49.21	27.71	53.99	37.60	47.27	59.10	33.03	42.14	35.99

Thank You!