StretchBEV: Stretching Future Instance Prediction Spatially and Temporally

Adil Kaan Akan, Fatma Güney
KUIS AI Center, Koç University

ECCV European Conference on Computer Vision, Tel Aviv 2022 — October 23-27, 2022, Tel Aviv

Future Instance Segmentation in BEV

Goal: Given a sequence of multi-camera images, predict the future instance segmentations in BEV.

Challenges:

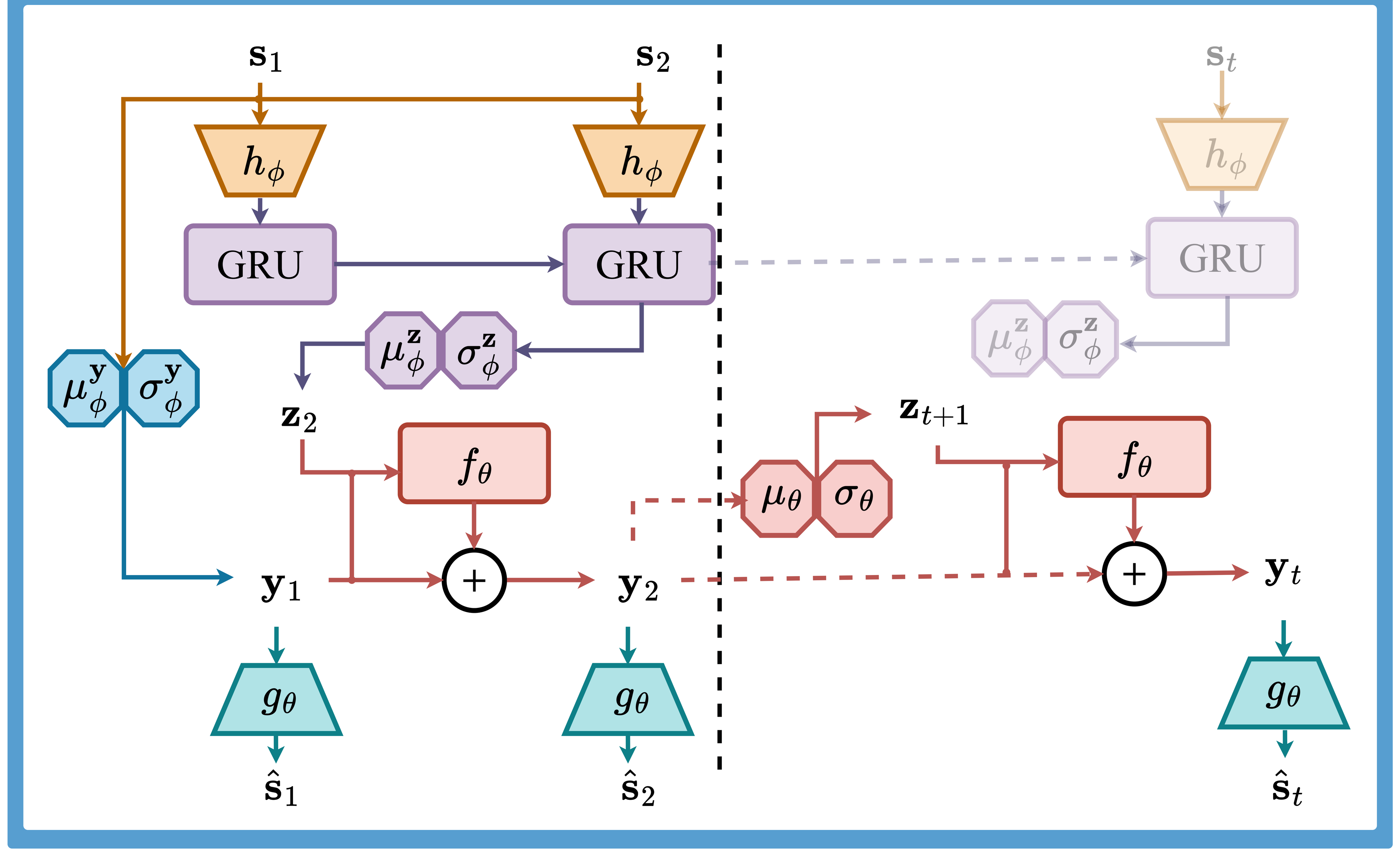
- High dimensionality of the problem
- Uncertainty of the future
- Requires an understanding of the scene structure, motion in the scene, and object relations, etc.

Input: Multi-camera sequences
Output: Future BEV instance segmentation

Inference

1. Encode the multi-camera images to create a BEV representation
2. Create first latent variable by_1 from the first 3 encoded states
3. Sample latent variables, bz_{t+1} from prior distributions
4. Predict the residual change in temporal dynamics and add it to the previous state
5. Decode predicted states to BEV states that are then decoded to instance segmentation

StretchBEV Model



Methodology

Time-Dependent Distributions

A separate latent variable for each time step t:

- $\mathbf{s}_t$: BEV state
- $\mathbf{y}_t$: Latent state variable
- $\mathbf{z}_t$: Stochastic latent variable
- $\mathbf{o}_t$: Output modalities (labels)

$$D_{\mathrm{KL}}(q(\mathbf{z}_t|\mathbf{s}_{1:t}, \mathbf{o}_{2:t}) \,||\, p(\mathbf{z}_t|\tilde{\mathbf{y}}_{t-1}))$$

vs.

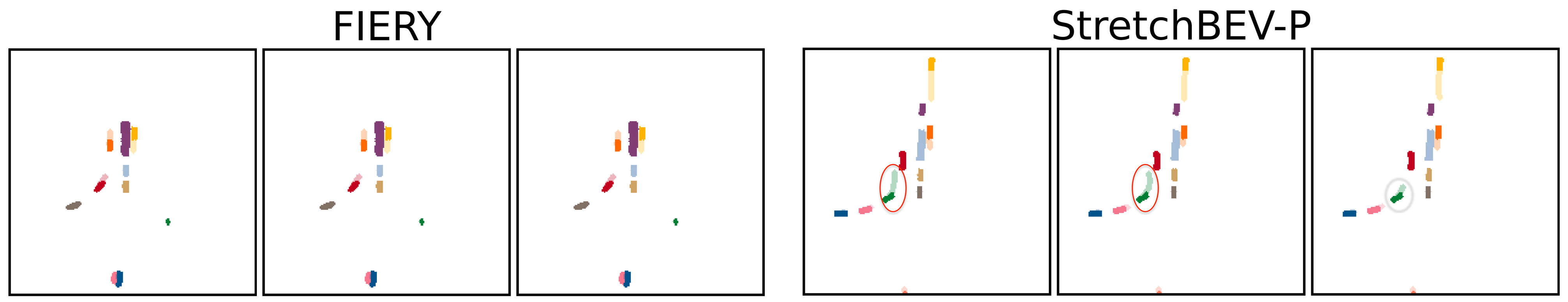
$$D_{\mathrm{KL}}(q(\mathbf{z}_{\text{future}}|\mathbf{s}_t, \mathbf{o}_{t+1:T}) \,||\, p(\mathbf{z}_{\text{present}}|\mathbf{s}_t))$$

Learning to Predict Future

1. Sample $\mathbf{z}_{t+1}$ from posterior containing future information $$\mathbf{z}_{t+1} \sim \mathcal{N}(\mu_\theta(\mathbf{y}_t), \sigma_\theta(\mathbf{y}_t)\ \mathbf{I})$$
2. Predict residual change to $\mathbf{y}_t$ based on sampled $\mathbf{z}_{t+1}$ $$\mathbf{y}_{t+1} = \mathbf{y}_t + f_\theta(\mathbf{y}_t, \mathbf{z}_{t+1})$$
3. Decode BEV state $\hat{\mathbf{s}}_t$ from $\mathbf{y}_t$ and $\hat{\mathbf{o}}_t$ from $\hat{\mathbf{s}}_t$ $$\hat{\mathbf{s}}_t = \mathbf{g}_\theta(\mathbf{y}_t)$$ $$\hat{\mathbf{o}}_t = \mathrm{Decoder}(\hat{\mathbf{s}}_t)$$

Qualitative Results

Qualitative Comparison to FIERY [1] on NuScenes



Diversity. Visualization of three samples

	Short				Mid				Long			
	IoU (↑)		VPQ (↑)		IoU (↑)		VPQ (↑)		IoU (↑)		VPQ (↑)	
	Near	Far	Near	Far	Near	Far	Near	Far	Near	Far	Near	Far
StretchBEV	55.5	37.1	46.0	29.0	**47.7**	32.5	39.1	23.8	**43.7**	28.4	36.4	21.0
FIERY	**58.8**	35.8	50.5	29.0	47.4	30.1	40.6	23.6	41.8	26.7	36.6	20.9
StretchBEV-P	58.1	**52.5**	**53.0**	**47.5**	46.8	**32.7**	**43.7**	**38.4**	38.2	**31.8**	**37.4**	**30.8**

Table: **Evaluation over Different Temporal Horizons.** Comparisons to FIERY [1] over short (2.0s), mid (4.0s), and long (6.0s) temporal horizons.

Conclusion

- Learning temporal dynamics through residual updates in the latent space
- SOTA and diverse results on all temporal horizons and regions

Contact Information

https://kuis-ai.github.io/stretchbev
{kakan20, fguney}@ku.edu.tr

References

[1] A. Hu, Z. Murez, N. Mohan, S. Dudas, J. Hawke, V. Badrinarayanan, R. Cipolla, and A. Kendall, "FIERY: Future instance segmentation in bird's-eye view from surround monocular cameras," 2021.