Paper ID: 8054

UL-VIO: Ultra-lightweight Visual-Inertial Odometry with Noise Robust Test-time Adaptation

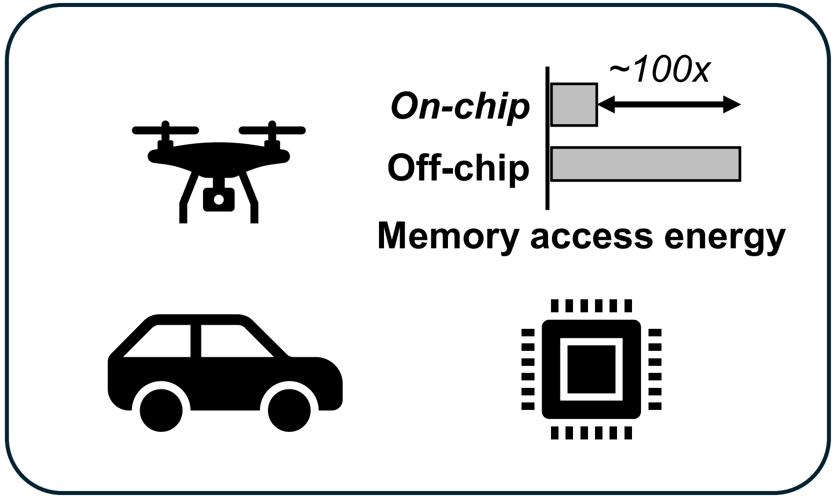
Jinho Park[1], Se Young Chun[2], and Mingoo Seok[1]
[1]Columbia University
[2]Seoul National University

jp4327@columbia.edu, sychun@snu.ac.kr, ms4415@columbia.edu

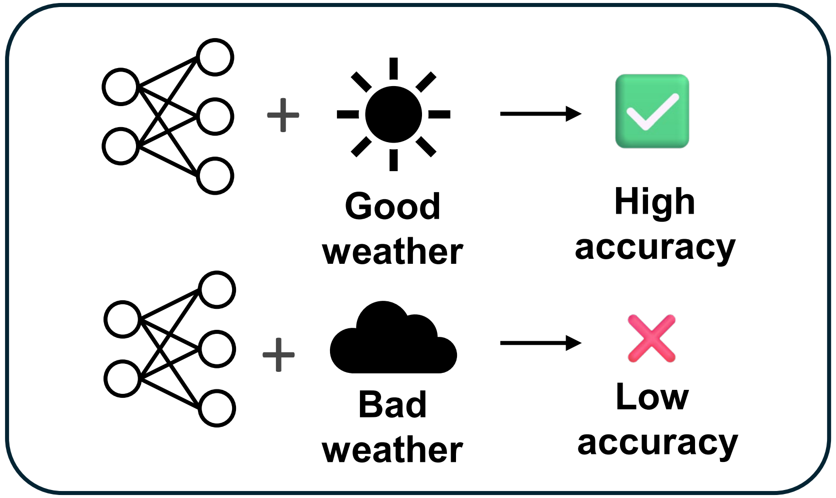
Table of Contents

VLSILab
@COLUMBIA UNIV

Motivation for Lightness & Robustness



Model compression

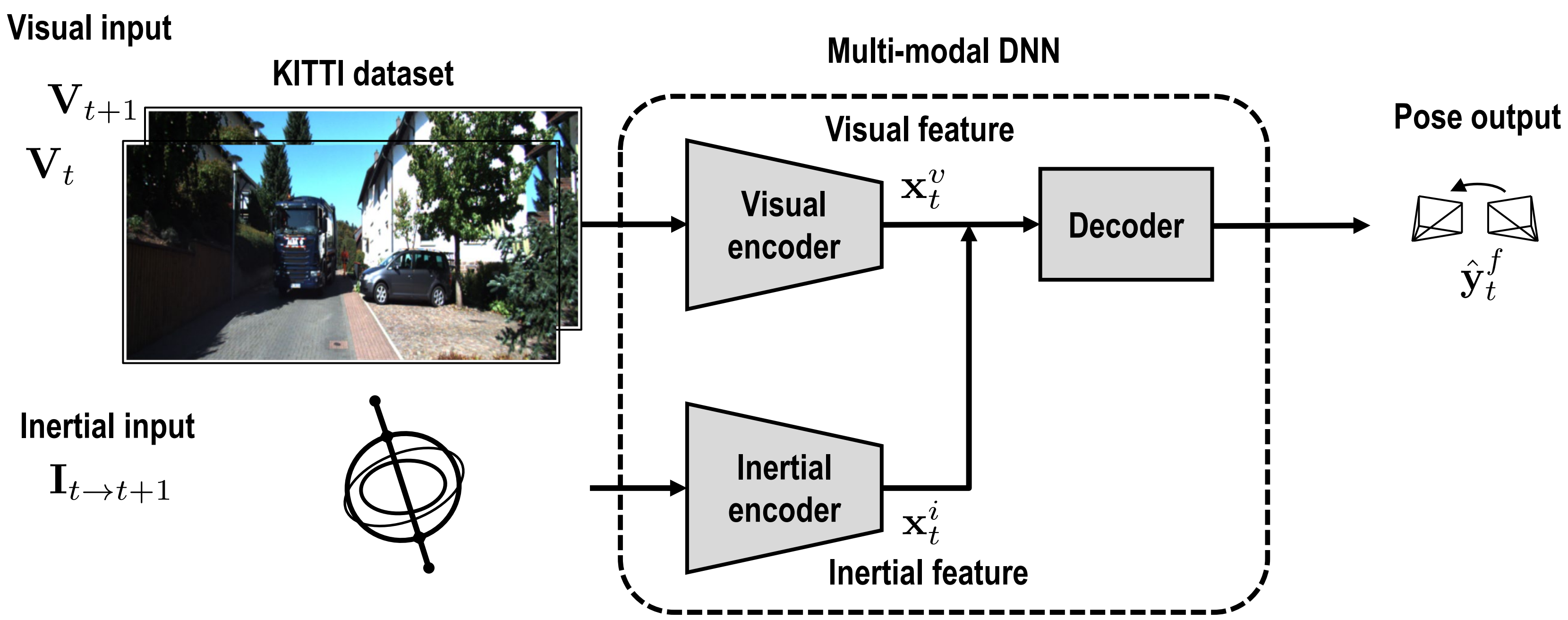


Test-time adaptation (TTA)

- Need for **lightweight** neural network to be hosted entirely by *on-chip* memory, which is of few MB in modern processors
- Robust to **environmental factors** and **noises** inducing distribution shifts

VLSILab @COLUMBIA UNIV

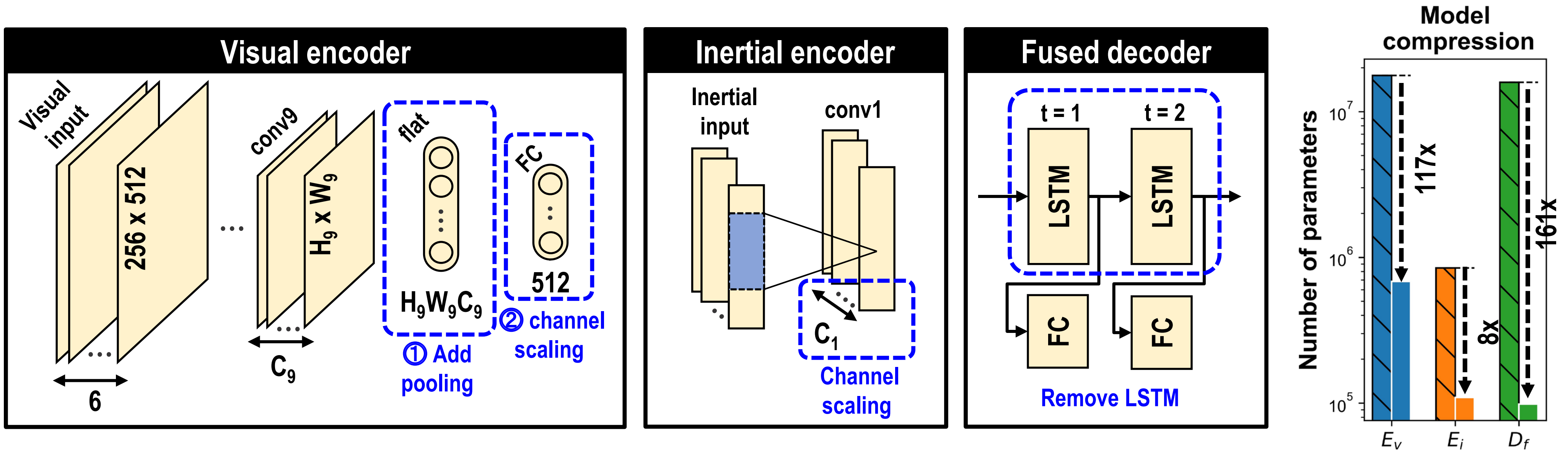
Learning-based Visual-inertial Odometry Pipeline



Given a sequence of visual and inertial inputs:

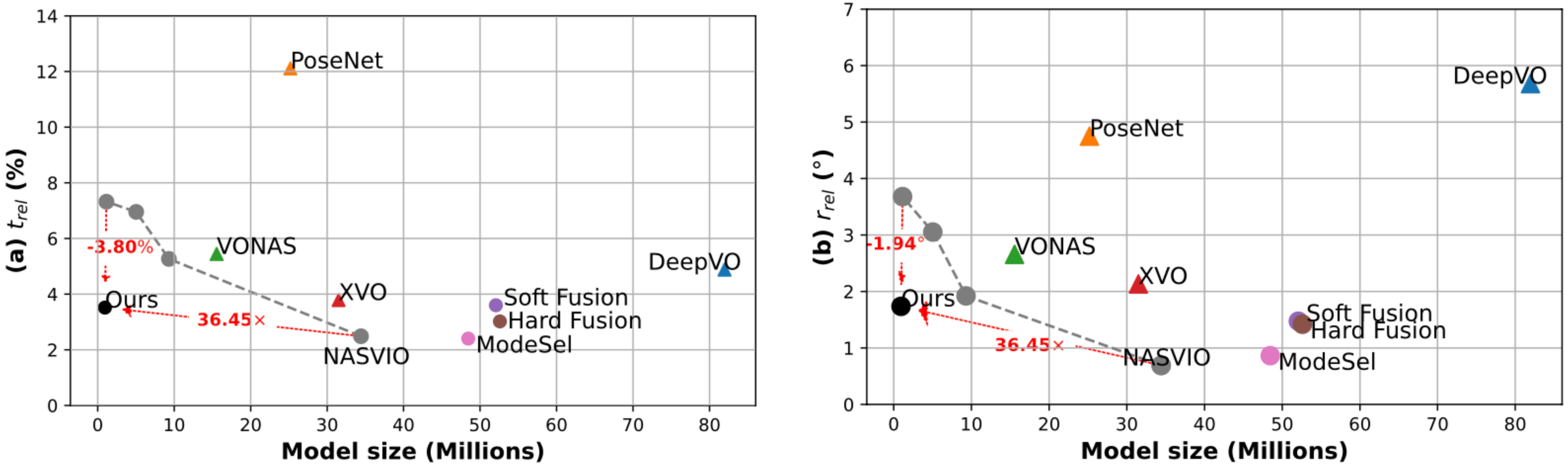
- **Visual encoder** E_v deduces visual features from pairwise images
- **Inertial encoder** E_i does so from the inertial input
- **Decoder** D predicts pose transformation from fused features

Model Compression



- Add an **AveragePool** after the last convolutional layer in E_v. This gives us 117x reduction in E_v
- Reduce the **channel size** in E_i since the parameter number is quadratically proportional to it, attaining 8x compression in E_i
- Replace the LSTM with **fully connected layers** for the D_f. This results in 161x downsizing in D_f

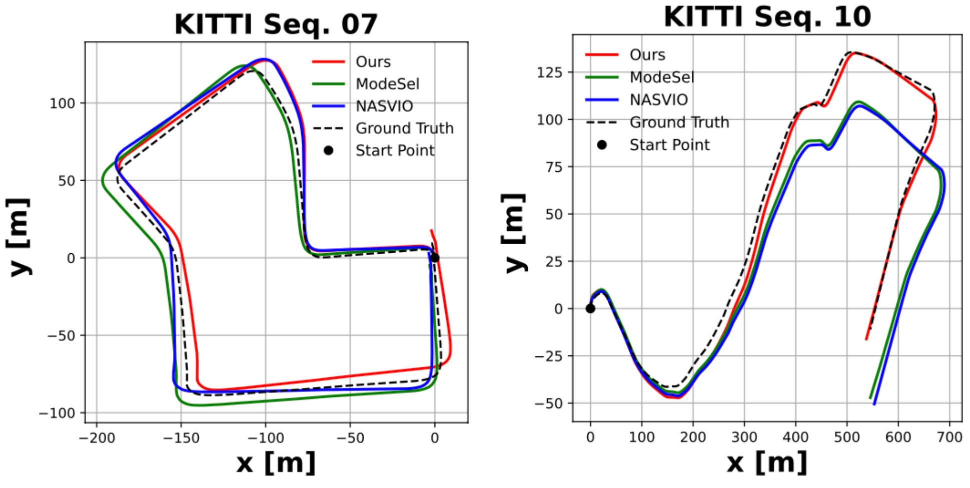
Model Compression – Results



- We achieve **36x model size reduction** w/ a minute (1%) increase in absolute pose estimation error
- Previous studies neglected **model size consideration** - major bottleneck in edge deployment
- Latest **Apple A16** and **Qualcomm Snapdragon** CPUs possess only a **few MB** of *on-chip* memory

VLSILab @COLUMBIA UNIV

Odometry Results

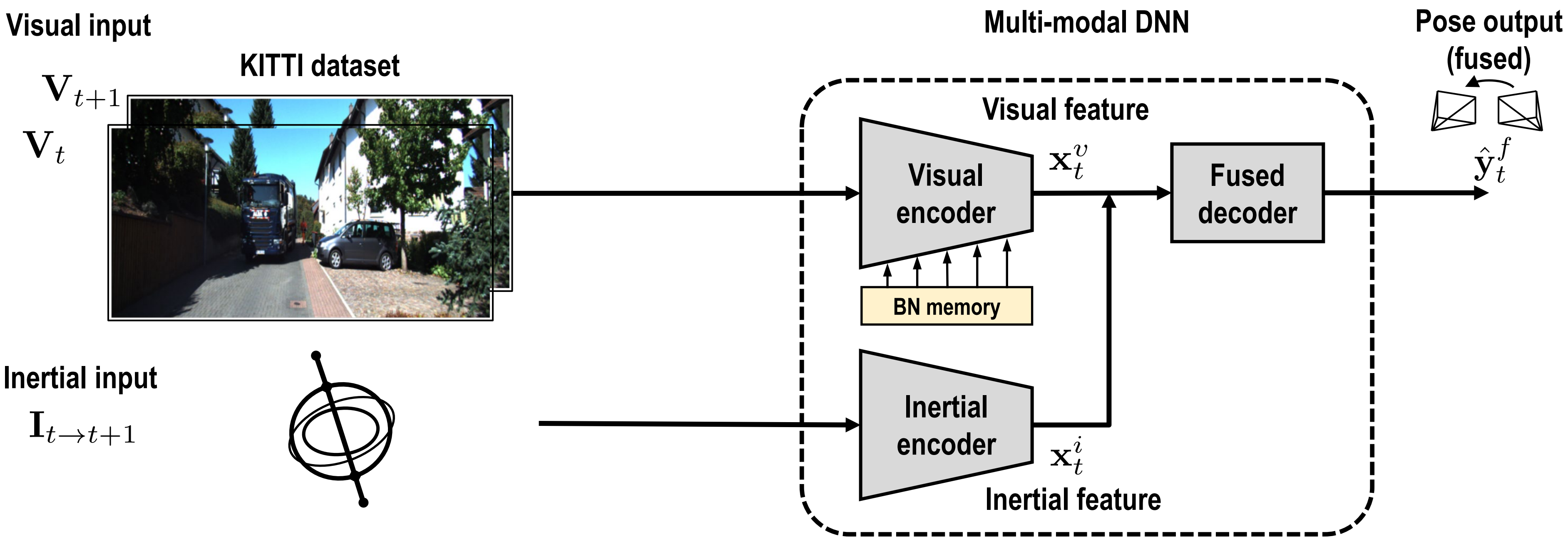


(a) (b)

	Ours	ModeSel [39]	Hard Fusion [5]
t_{rmse} [m]	0.0282	**0.0178** (−0.0104)	0.0283 (+0.0001)
r_{rmse} (°)	0.0756	0.0906 (+0.0150)	**0.0402** (−0.0354)
Model size (M)	**0.944**	48.454 (×51.3)	52.598 (×55.7)

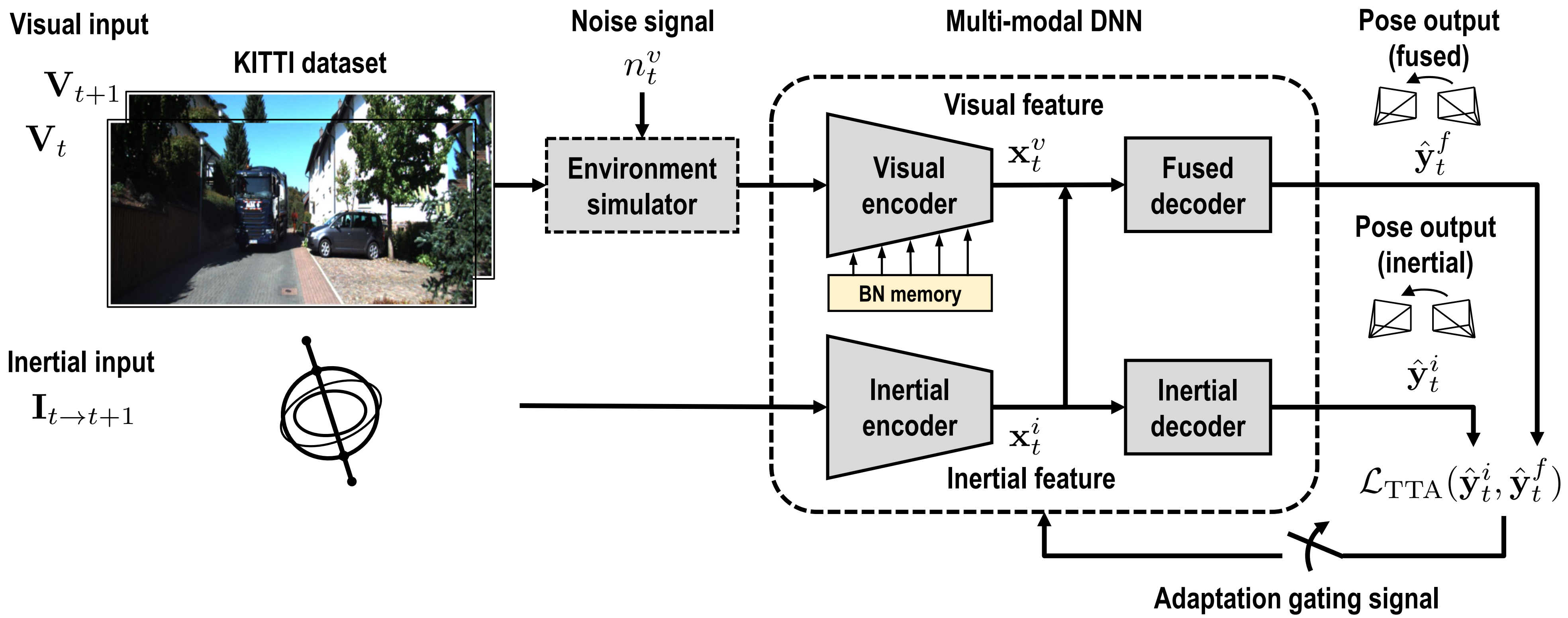
- In **KITTI,** ours with model compression has a slight deviation Seq. 07 and better ego-motion in Seq. 10 than ModeSel and NASVIO
- In **EuRoC** dataset, ours perform comparably to ModeSel [ECCV`22] and Hard Fusion [CVPR`19] with >50X reduction.

Proposed TTA w/ Multi-modal Consistency



Keep the original VIO network intact:

Proposed TTA w/ Multi-modal Consistency

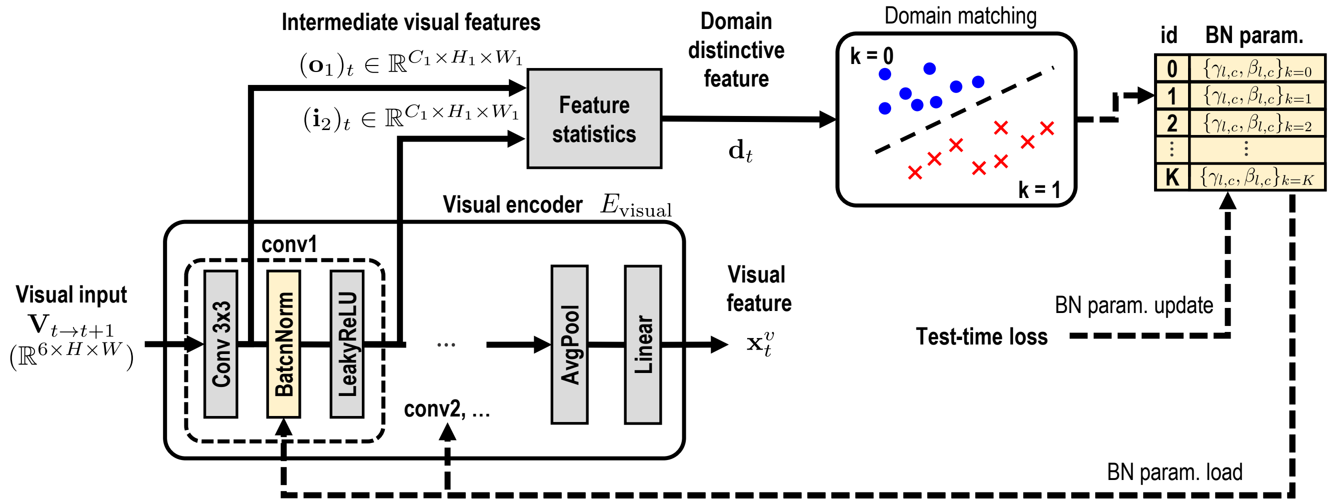


Keep the original VIO network intact:

- Visual inputs now experience **noise** from the **environment**
- Dedicate a separate **inertial decoder** using just the inertial feature
- **Adapt** the network based on inertial only feature at **test-time**

VLSILab @COLUMBIA UNIV

Dictionary-based Adaptation



Algorithm 1: Online TTA with adaptation gating

Input: Camera sequence $(\{\mathbf{V}_t\}_{t=1}^{T})$, IMU sequence $(\{\mathbf{I}_t\}_{t=1}^{T})$, frozen weight $(\mathbf{\Theta}_f)$, adaptation weight $(\{\mathbf{\Theta}_a^k\}_{k=0}^{K})$, domain distinctive feature $(\{\mathbf{d}^k\}_{k=0}^{K})$, learning rate (η)

Output: Pose transformation sequence $(\{\hat{\mathbf{y}}^t\}_{t=1}^{T-1})$

1: **for** $t := 1$ *to* $T-1$ **do**

2: $\hat{\mathbf{y}}_f, \hat{\mathbf{y}}_i, \hat{\mathbf{d}}_t \leftarrow f(\mathbf{V}_{t\to t+1}, \mathbf{I}_{t\to t+1}, \mathbf{\Theta}_k)$;

3: $k \leftarrow \text{Match}(\hat{\mathbf{d}}_t, \mathbf{d}^k)$; // Eq. 6

4: **if** $k \neq 0$ **then**

5: $\mathbf{\Theta}_a^k \leftarrow \mathbf{\Theta}_a^k - \eta \nabla_{\mathbf{\Theta}} \mathcal{L}_{\text{TTA}}(\hat{\mathbf{y}}_f, \hat{\mathbf{y}}_i)$; // BatchNorm parameter update

Domain distinctive feature (*ddf*) $\hat{\mathbf{d}} = \mu(\mathbf{o}_1) \| \sigma(\mathbf{o}_1) \| \mu(\mathbf{i}_2) \| \sigma(\mathbf{i}_2)$

l_2 distance-based domain search $k_t = \underset{k\in[0,1,..,K]}{\arg\min} \|\hat{\mathbf{d}}_t - \mathbf{d}^k\|_2$

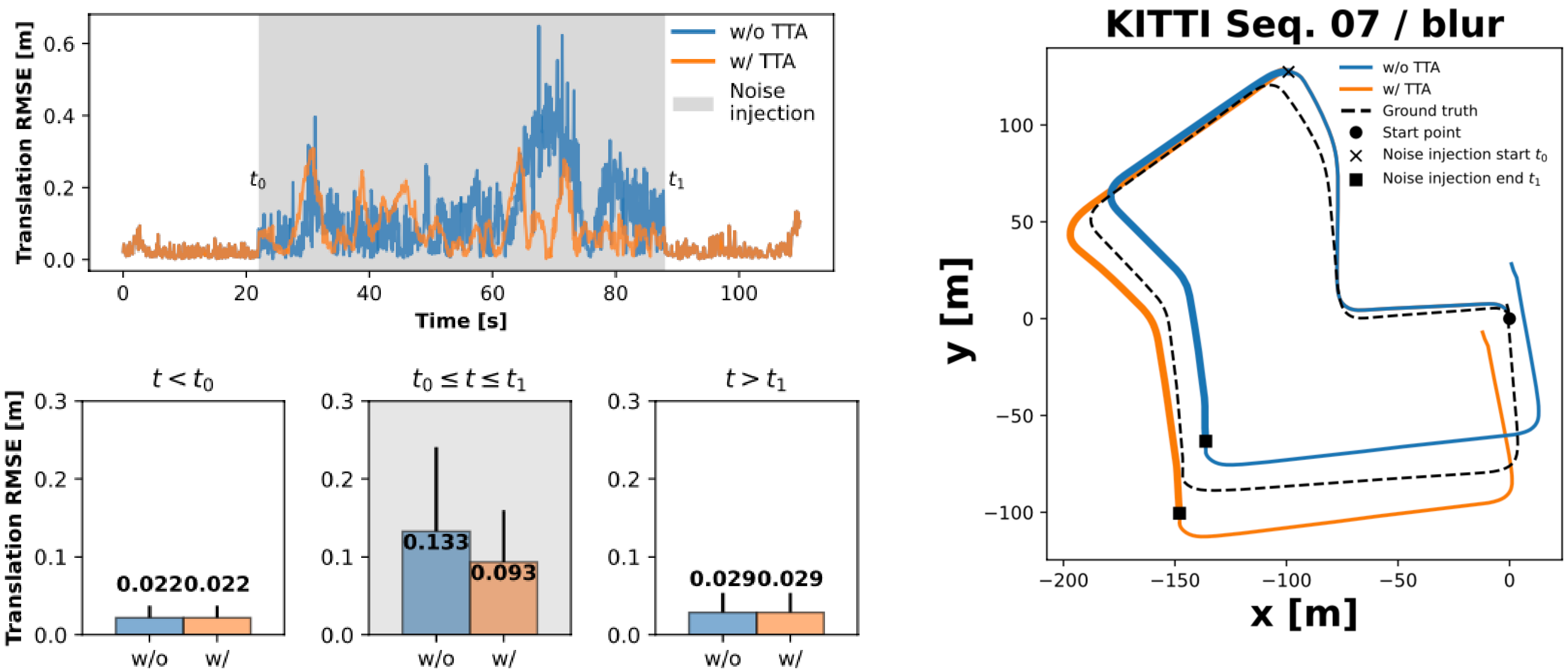
- Forward-path generate **visual feature** x^v, which will be used for decoder in the deeper layers (Line 2)
- First layer conv output $\mathbf{o}_1$ is used to generate **domain distinctive feature** $\mathbf{d}_t$ (Line 2)
- Domain k is found using l_2 distance with the proxies $\{\mathbf{d}^k\}$ (Line 3)
- **BatchNorm** parameter corresponding to the domain k is updated (Line 5)

Comparison w/ Fine-tuned Baselines

Model		Clean	Multi.	Blur	Rain	Snow	Shadow	Bright.	Cont.
		Average pose-wise t_{rmse} [m]							
Source		0.059	0.154	<u>0.261</u>	0.176	0.191	0.203	0.226	<u>0.250</u>
Fine-tuned with adver. noise (FT)	Multi.	0.099	<u>0.129</u>	0.394	0.227	0.372	0.192	0.299	0.331
	Blur	0.115	0.176	0.263	0.193	0.247	0.184	0.242	0.261
	Rain	0.289	0.325	0.372	**0.095**	0.394	0.311	0.525	0.531
	Snow	0.091	0.148	0.319	0.263	<u>0.183</u>	0.208	0.369	0.450
	Shadow	0.085	**0.112**	0.322	0.179	0.243	**0.121**	<u>0.221</u>	0.252
	Bright.	0.091	0.151	0.312	0.177	0.226	0.185	0.233	0.278
	Cont.	0.093	0.150	0.330	0.197	0.219	0.184	0.237	0.273
TTA (ours)		-	0.156	**0.230**	<u>0.143</u>	**0.172**	<u>0.155</u>	**0.193**	**0.212**

- We demonstrate the effectiveness of our **TTA** method by comparing it with networks **fine-tuned** with adversarial noises
- Except for one case, e.g., multiplicative noise, our TTA method has **the best or second-best accuracy**

Online TTA & Trajectory Results



- Translation RMSE quickly increases when the **visual corruption** is applied
- **TTA** mitigates the increase in the error to some extent
- Noise injected causes the trajectory to be **underestimated**, which is alleviated w/ TTA

VLSILab @COLUMBIA UNIV

Continual TTA

KITTI

Seq. / Noise	Seq. 05 Blur	Seq. 05 Rain	Seq. 05 Snow	Seq. 05 Con.	Seq. 07 Blur	Seq. 07 Rain	Seq. 07 Snow	Seq. 07 Con.	Seq. 10 Blur	Seq. 10 Rain	Seq. 10 Snow	Seq. 10 Con.	Avg.
Baseline	0.118	0.121	**0.103**	0.166	0.127	0.153	0.110	0.191	0.137	0.134	**0.120**	0.167	0.137
TTA	**0.112**	**0.107**	0.110	**0.107**	**0.101**	**0.108**	**0.106**	**0.104**	**0.123**	**0.124**	0.121	**0.127**	**0.113**
***ddf* acc.**	97.9	100	100	100	98.2	100	100	100	98.8	100	100	100	99.6

Time t →

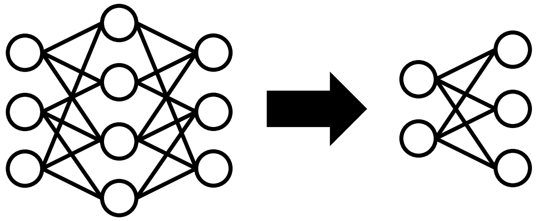
EuRoC

Noise	Blur	Bright.	Contrast	Avg.
Baseline	0.0255	0.0256	0.0276	0.0262
TTA	**0.0253**	**0.0254**	**0.0254**	**0.0254**
***ddf* acc. (%)**	95.6	100.0	100.0	98.5

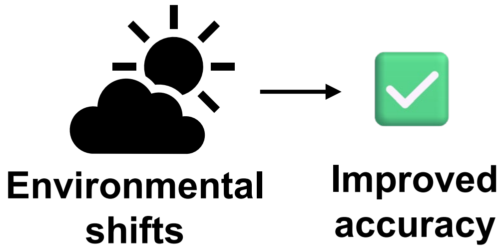
Time t →

- We inject different types of noise {blur, rain, snow, contrast} in a continuous manner for **KITTI**
- Visual corruption {blur, brightness, contrast} relatable to indoor for **EuRoC**
- TTA effectively reduces that translation RMSE

Conclusion



Model compression



Test-time adaptation

- We achieve a **network** with **< 1M parameter size** through model compression, *i.e.* 36x less than SoTA

- We effectively alleviate pose estimation error with multi-modal consistency-based **TTA**

VLSILab
@COLUMBIA UNIV

Thank you for listening to our work on UL-VIO:

Ultra-lightweight Visual-Inertial Odometry with Noise Robust Test-time Adaptation

Project page



Jinho Park[1] **Se Young Chun**[2] **Mingoo Seok**[1]

[1]Columbia University
[2]Seoul National University

jp4327@columbia.edu, sychun@snu.ac.kr, ms4415@columbia.edu

Acknowledgements:
This work was supported in part by COGNISENSE, one of seven centers in JUMP 2.0, a Semiconductor Research Corporation (SRC) program sponsored by DARPA. The work of SY Chun was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)] and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) [No. NRF-2022M3C1A309202211]

VLSILab @COLUMBIA UNIV