Valeo

Making sense of models that know (too) much

Reliability in the age of Foundation Models

Andrei BURSUC
valeo.ai
12 September 2025

Valeo's history in ADAS

1991 – Ultrasonic Parking Aid

2004 – Lane-Departure Warning

2006 – Multi-beam Radar

2007 – Semi-automatic Park Assist

2013 – Object detection in Rear Camera

2016 – Remote Park Assist

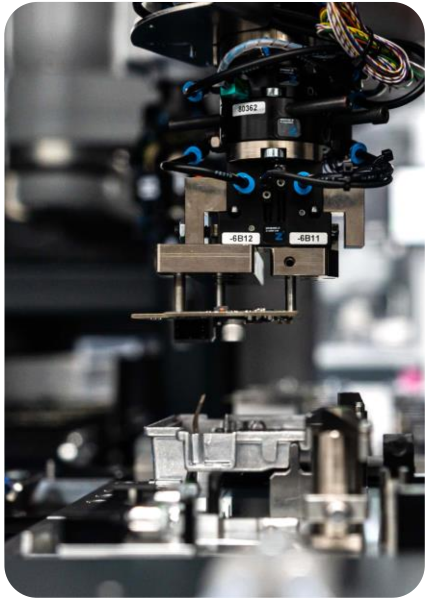
2020 – Vision-only L2 System

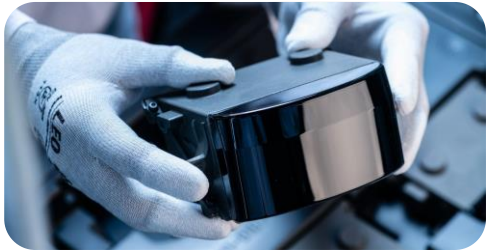
2022 – SCALA 2 for 1st consumer L3

2023 – Complete L2 ADAS System

2025 – ADAS Domain Controller

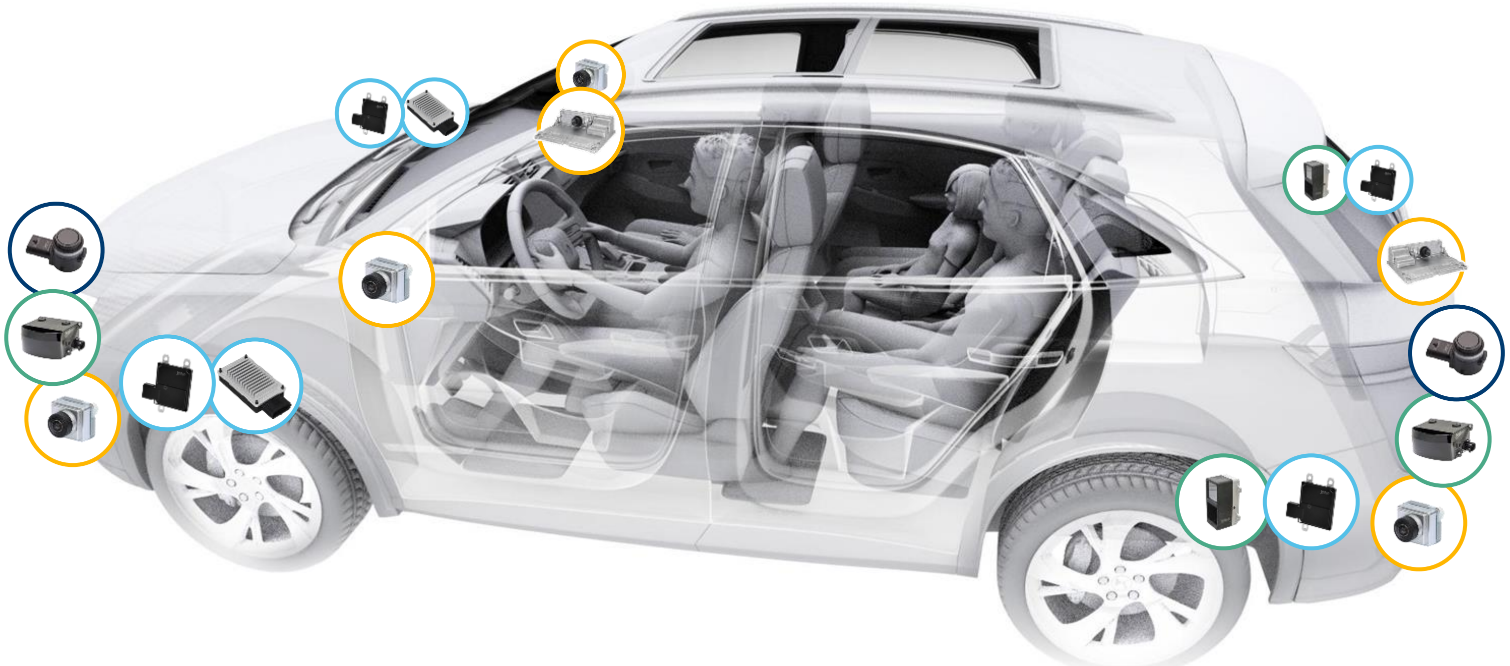
1.5+ Billion sensors shipped in 30 years





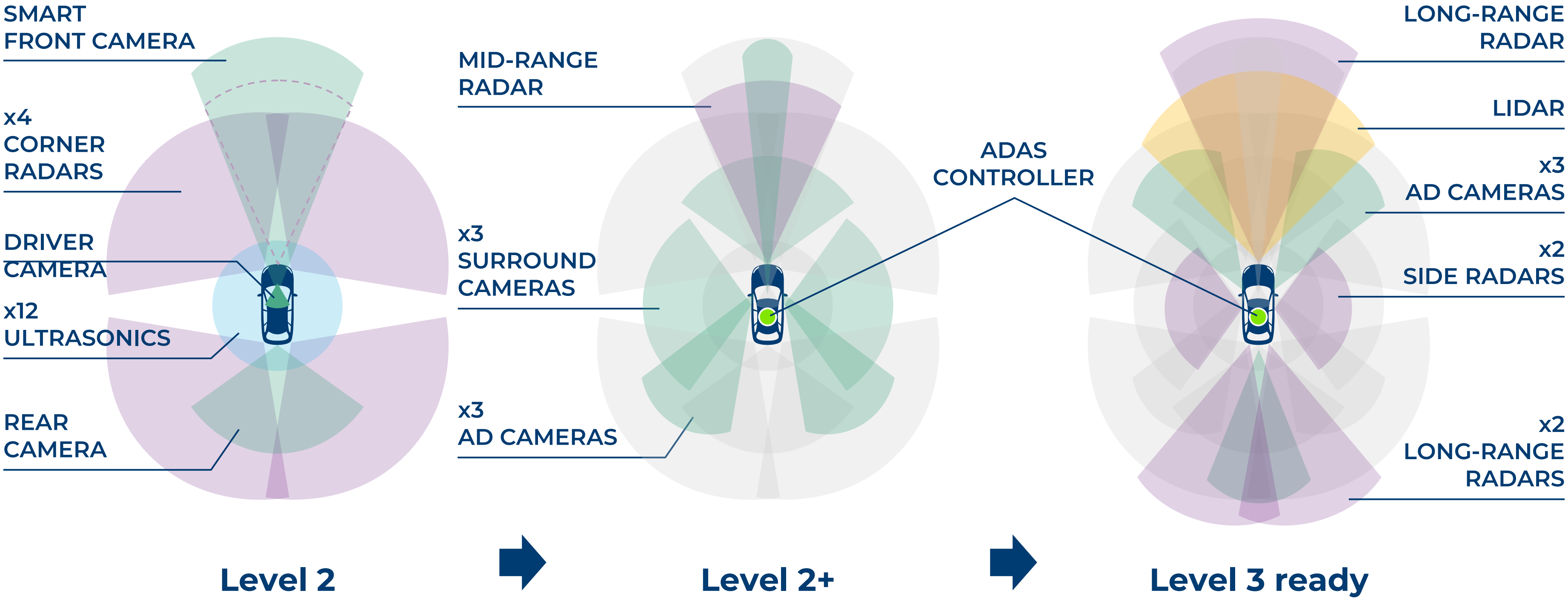
Another **1.5+ billion sensors** to be shipped in the next 5 years

Valeo sensor suite



ULS	RADARS		CAMERAS		LIDARS	
Ultrasonic sensors	Near field radars	Mid range radars	Surround view cameras	Long range cameras	Near field lidars	SCALA

Scalable system architecture



From ADAS* to AD**

Spectrum of vehicle automatization

Driving Assistance

- Blind spot detection
- Cruise control

Forward collision warning
+
autobrake

↓ **56%** Front-to-rear crashes with injuries

Lane departure warning

↓ **21%** Injury crashes

*ADAS = Advanced Driving Assistance Systems **AD = Autonomous driving

From ADAS* to AD**

Spectrum of vehicle automatization

Driving Assistance

- Blind spot detection
- Cruise control

Limited Self-Driving

- Parking valet
- Highway pilot

Full Self-Driving

- Robot taxis
- Delivery vehicle

Towards safer, more efficient and more available mobility

*ADAS = Advanced Driving Assistance Systems **AD = Autonomous driving

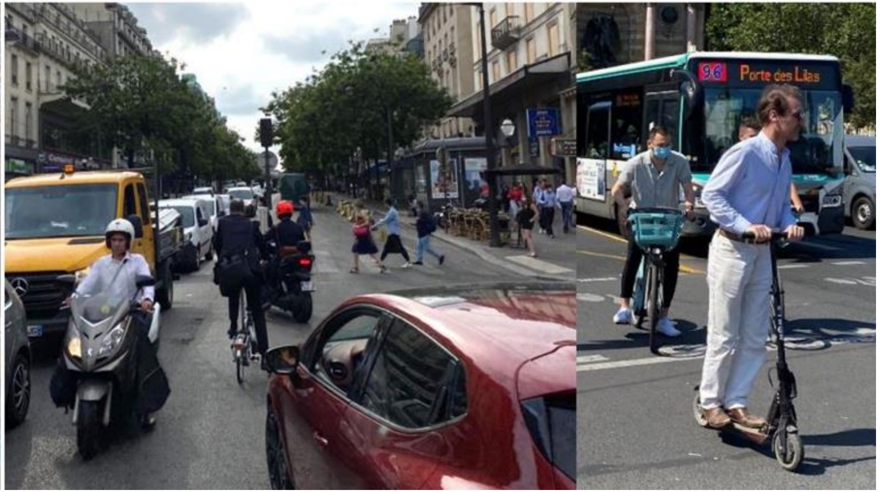
valeo.ai

- ~25 researchers & PhDs
- Dedicated to open research
- 10s of academic collabs across France and Europe
- Offices: Paris, Prague
- Topics: perception, data efficiency, forecasting, reliability, explainability

https://valeoai.github.io

Hello world!


Café Journal
cash
Holborn
Liverpool Street 8
Bow Church



96 Porte des Lilas

1.36M

deaths due to vehicle crashes each year

42,915

deaths in the U.S. in 2021 and 2.5 million injuries

$836B

in harm from loss of life and injury each year

50M

Injuries worldwide due to vehicle crashes each year

79%

of seniors age 65 and older live in car-dependent communities.

12M

people 40 years and over in the United States have vision impairment.

From intended to covered domain

Dataset defines the actual domain, often with limited coverage of:

- Rare pose/appearance of known objects, rare objects
- Rare, e.g., dangerous, scene configurations
- All sorts of perturbation, e.g., adverse conditions, sensor blocking








Uncertainty estimation

Good uncertainty estimates quantify when we can trust the model's predictions → helps avoid mistakes or select difficult data to be labelled.

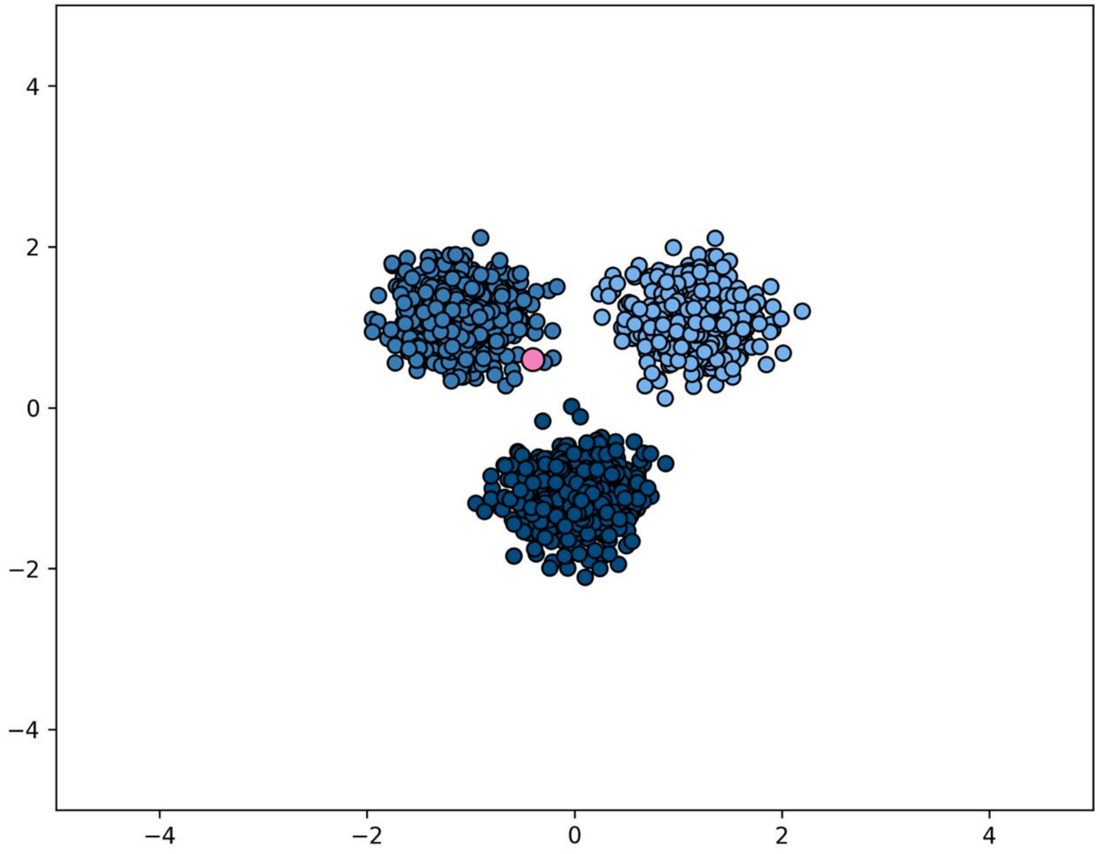
Uncertainty estimation is an essential function for improving reliability and safety of systems running on ML models.

Sources of uncertainty

(quick recap)

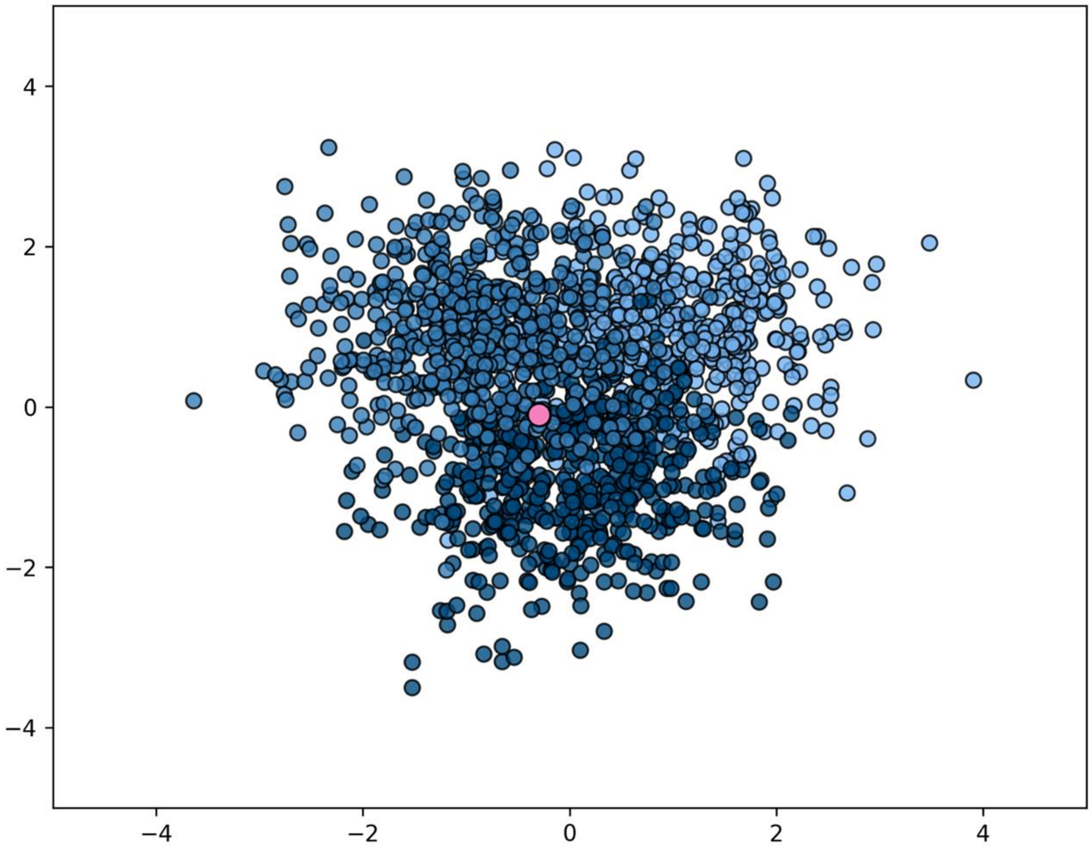
There are two main types of uncertainties each with its own peculiarities

Case 1



A. Malinin, Uncertainty Estimation in Deep Learning with application to Spoken Language Assessment, PhD Thesis 2019

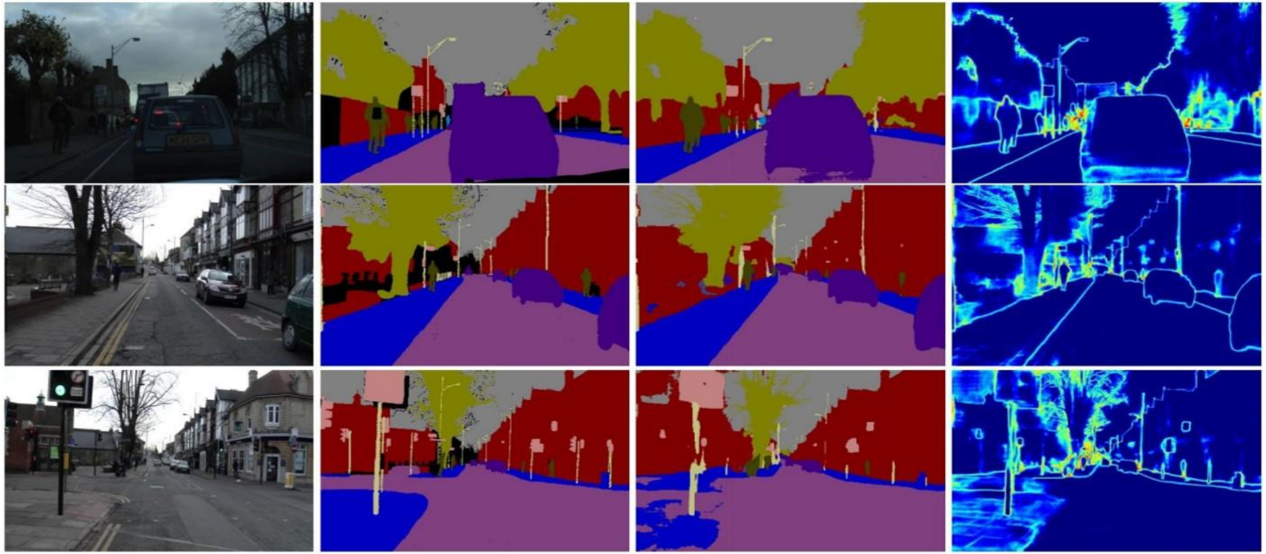
Data / Aleatoric uncertainty



A. Malinin, Uncertainty Estimation in Deep Learning with application to Spoken Language Assessment, PhD Thesis 2019

Data / Aleatoric uncertainty

Similarly looking objects also fall into this category.

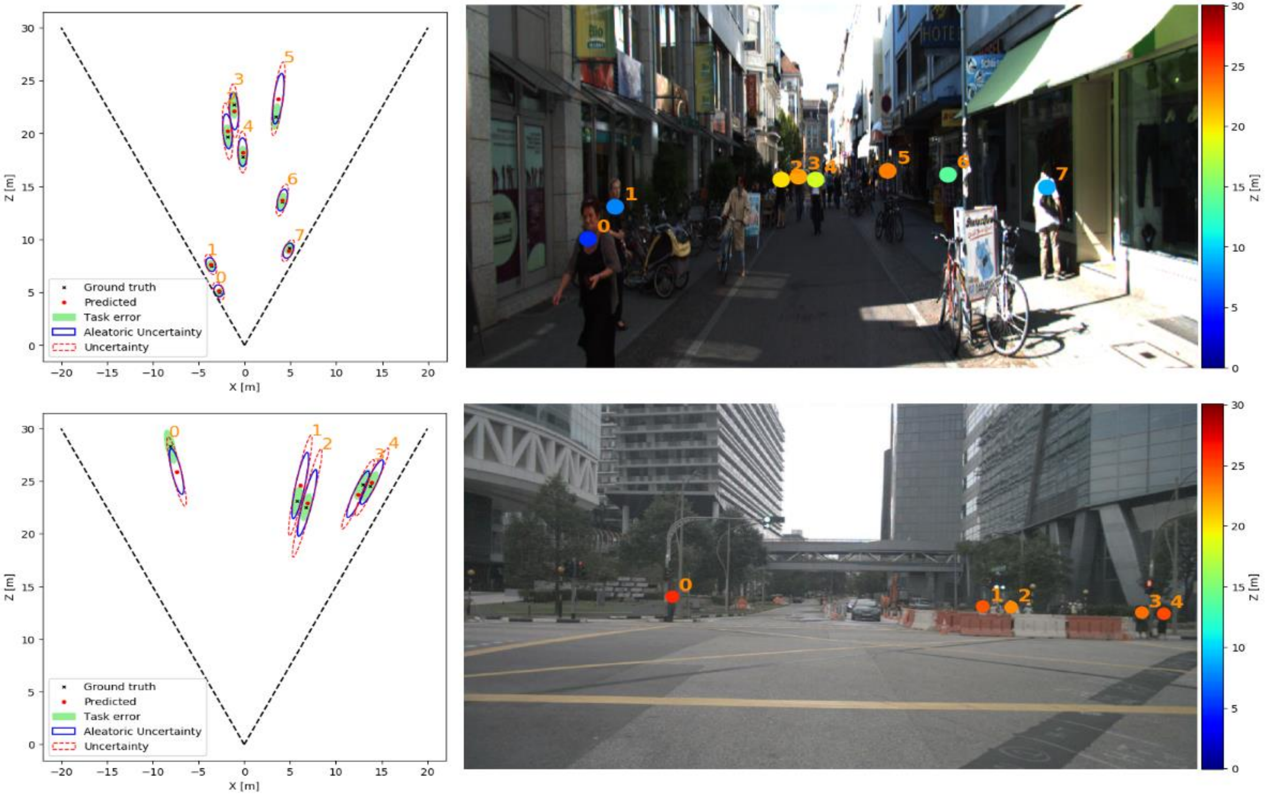


(a) Input Image (b) Ground Truth (c) Semantic Segmentation (d) Aleatoric Uncertainty

In urban scenes this type of uncertainty is frequently caused by similarly-looking classes:

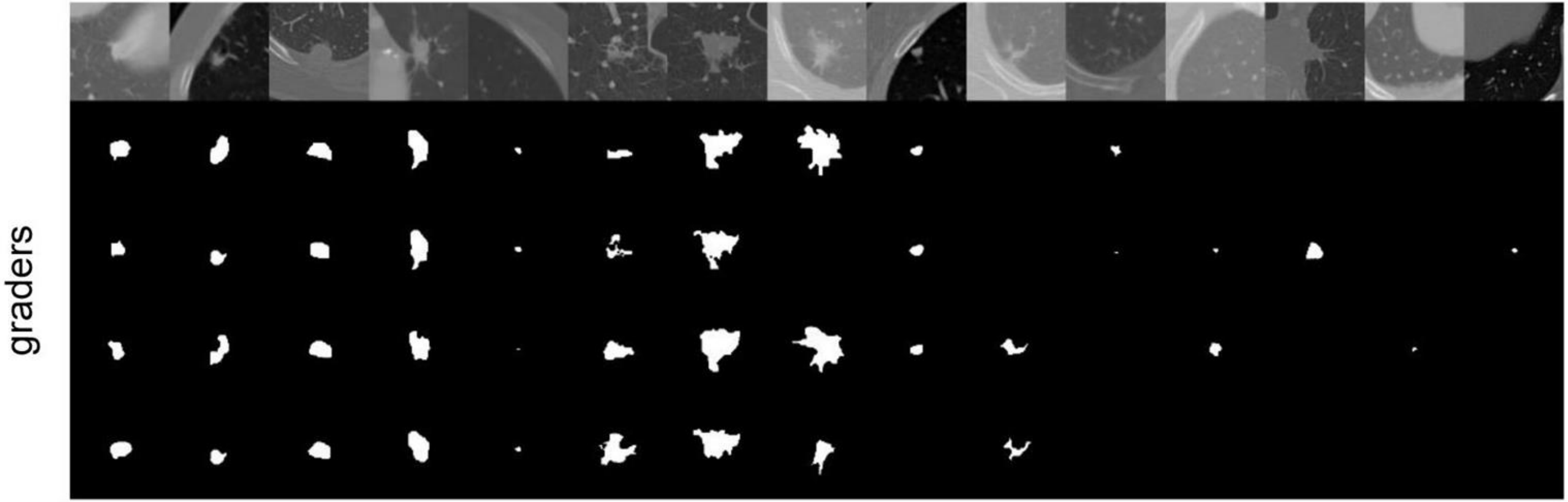
- pedestrian - cyclist - person on trottinette/scooter
- road - sidewalk
- also at object boundaries

A. Kendall and Y. Gal, What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?, NeurIPS 2017



Also caused by sensor limitations: localization and recognition of far-away objects is less precise. Datasets with low resolution images, e.g., CIFAR, also expose this ambiguity.

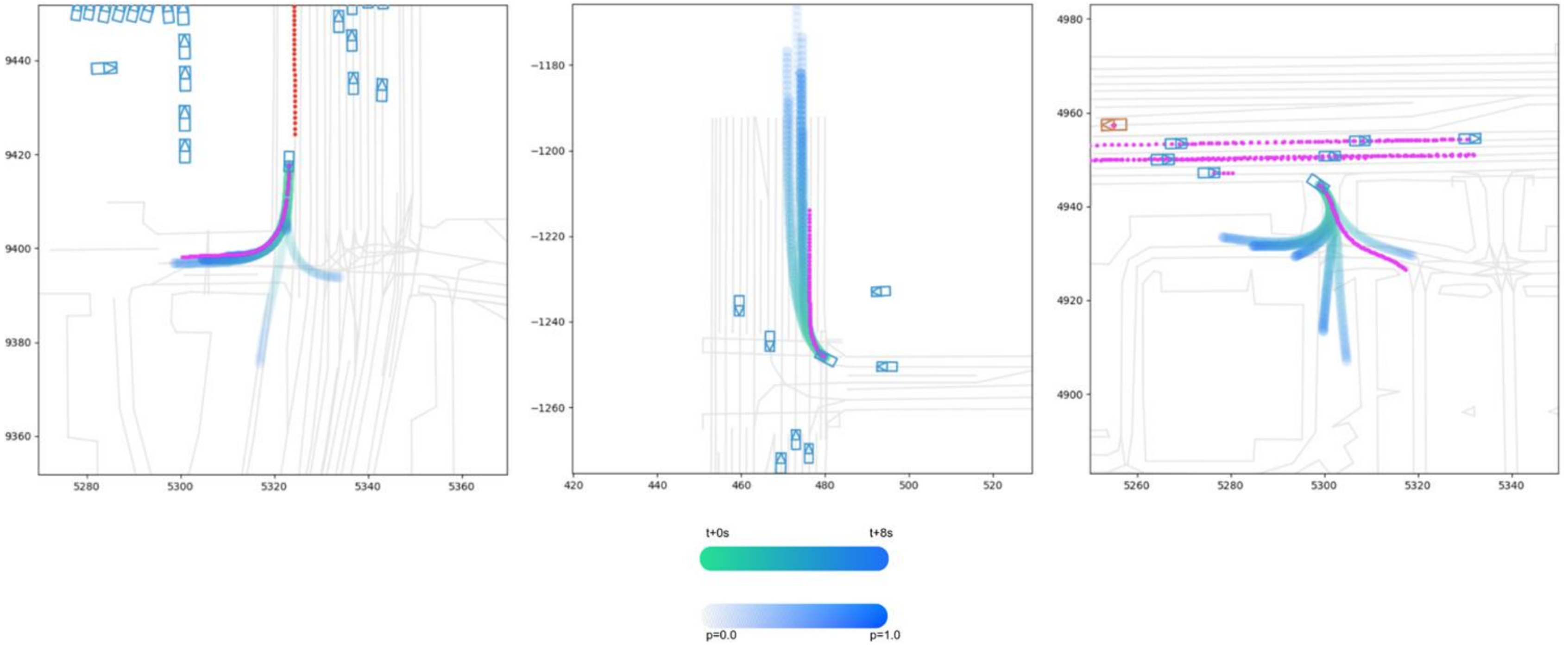
L. Bertoni et al., MonoLoco: Monocular 3D Pedestrian Localization and Uncertainty Estimation, ICCV 2019



Samples and annotations from different graders on LIDC-IDRI dataset.

Difficult or ambiguous samples with annotation disagreement

S.G. Armato et al., The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans, Medical Physics 2011
S. Kohl et al., A Probabilistic U-Net for Segmentation of Ambiguous Images, NeurIPS 2018



Multiple potential outcomes - motion forecasting

B. Varadarajan et al., MultiPath++: Efficient Information Fusion and Trajectory Aggregation for Behavior Prediction, ICRA 2022

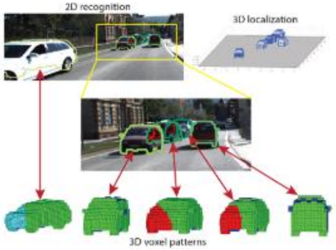
Data uncertainty

*Rain drops**

Lack of visual features

Glare

Low light



Occlusion

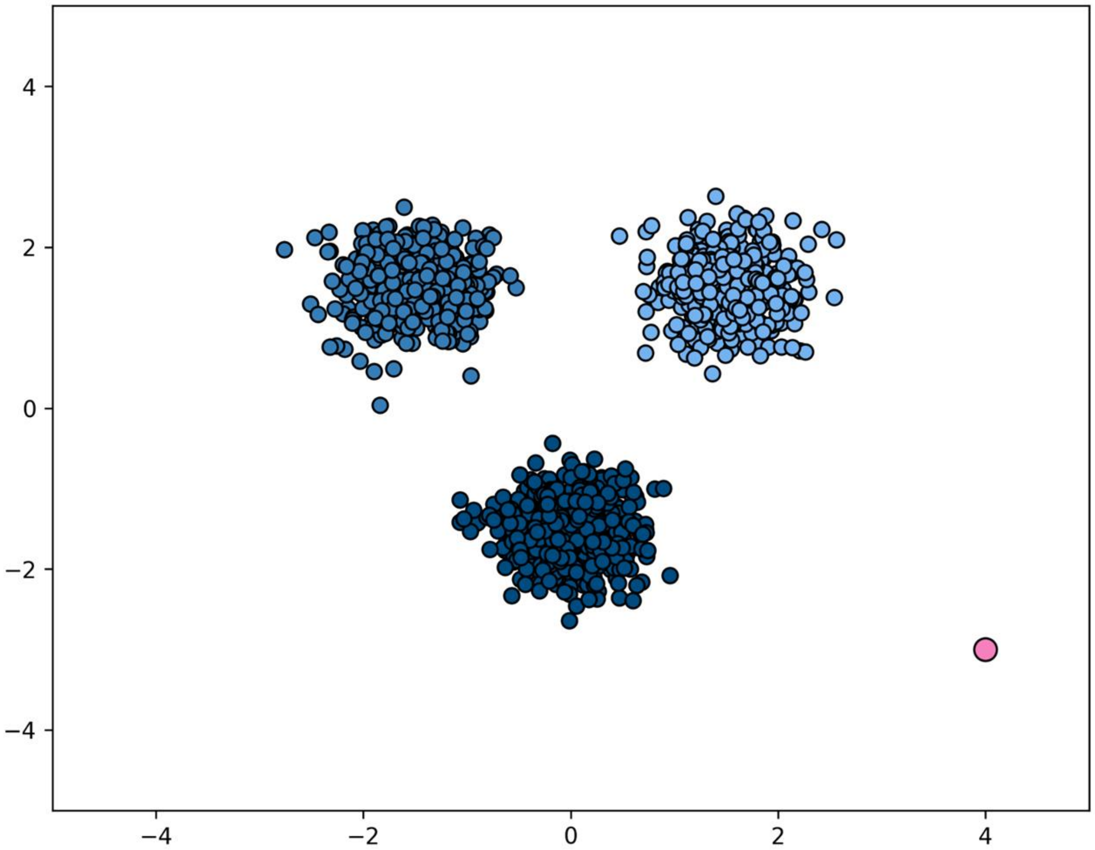
- Data uncertainty is often encountered in practice due to sensor quality, natural randomness, that cannot be explained by our data.

- Uncertainty due to the properties of the data

- It cannot be reduced (irreducible uncertainty), but can be learned. Could be reduced with better measurements.

Case 2

Knowledge / Epistemic uncertainty



A. Malinin, Uncertainty Estimation in Deep Learning with application to Spoken Language Assessment, PhD Thesis 2019

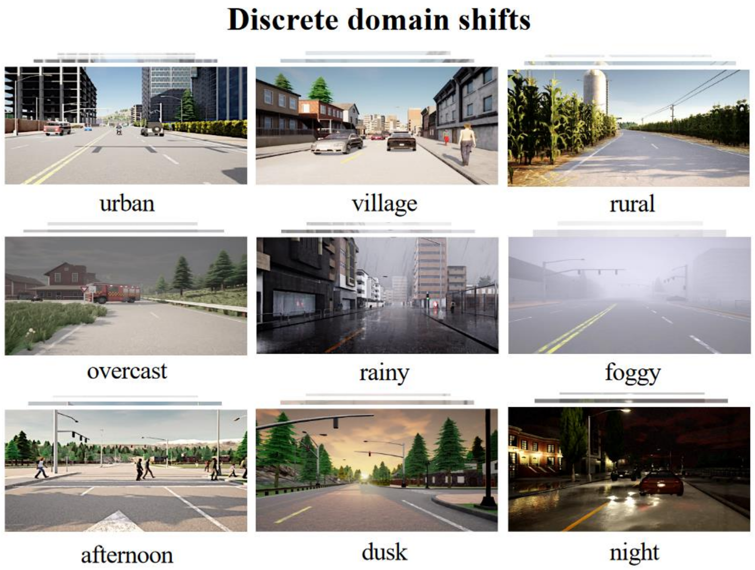
Data uncertainty

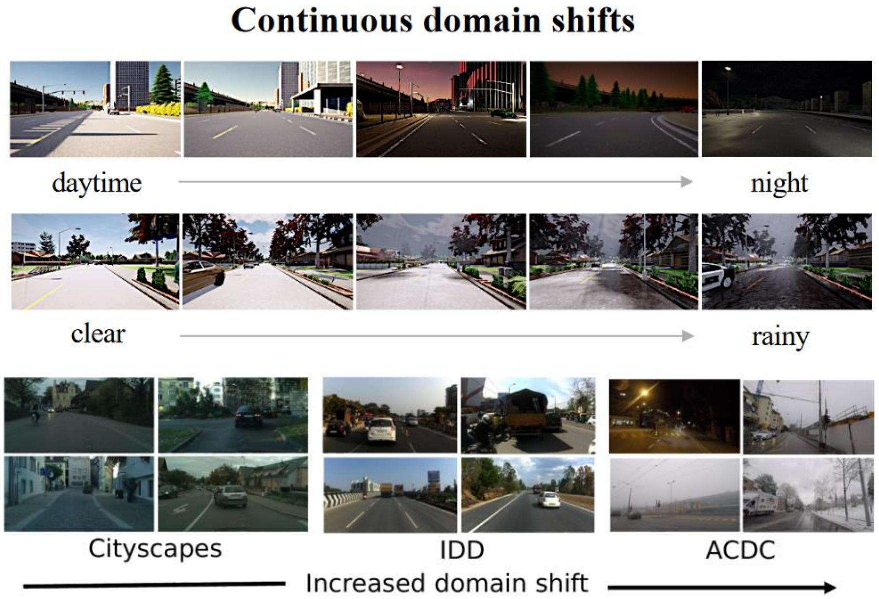
I.I.D.: $p_{\text{train}}(x, y) = p_{\text{test}}(x, y)$ **O.O.D.:** $p_{\text{train}}(x, y) \neq p_{\text{test}}(x, y)$

There are different forms of out-of-distribution / distribution shift:

- covariate shift: distribution of $p(x)$ changes, while $p(y \mid x)$ remains constant
- label shift: distribution of labels $p(y)$: changes, while $p(x \mid y)$ remains constant
- OOD or anomaly: new object classes appear at test time

Domain shift





Distribution shift of varying degrees is often encountered in real world conditions

---

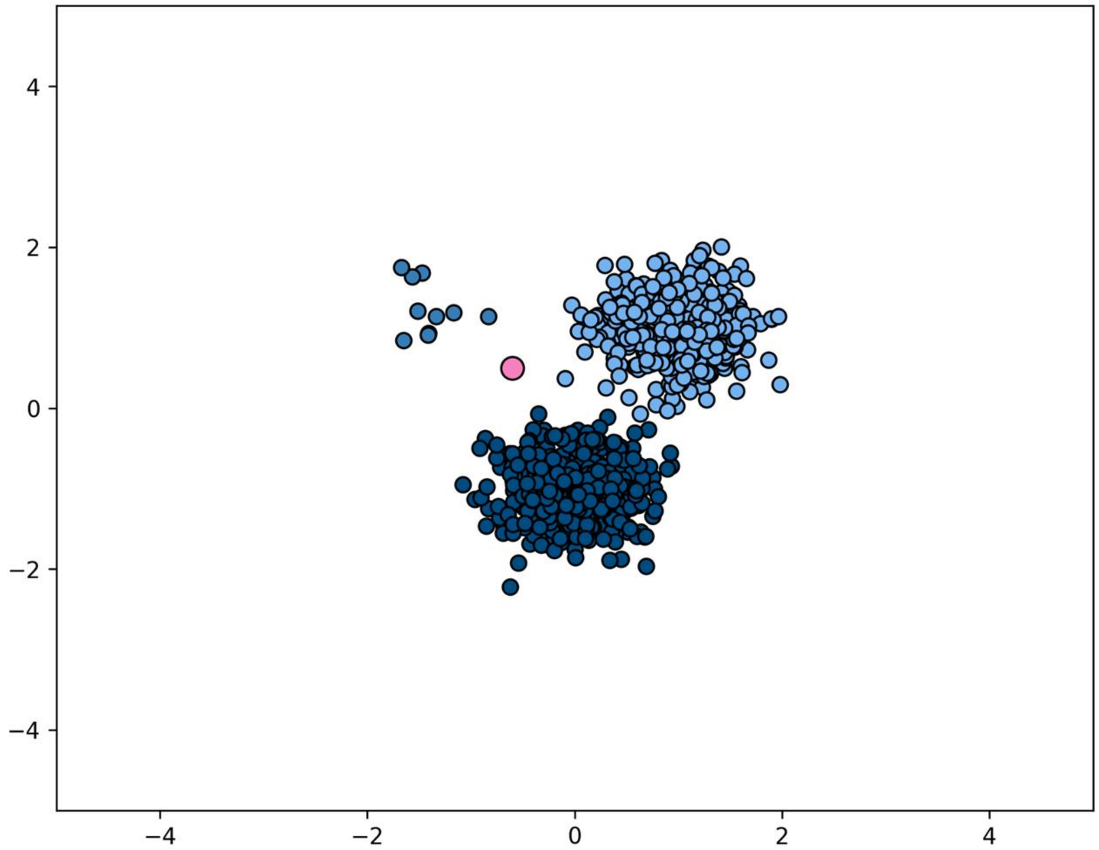
T. Sun et al., SHIFT: A Synthetic Driving Dataset for Continuous Multi-Task Domain Adaptation, CVPR 2022
P. de Jorge et al., Reliability in Semantic Segmentation: Are We on the Right Track?, CVPR 2023

Object-level shift

Row 1: Lost&Found; Row 2: SegmentMeIfYoucan; Row 3: BRAVO synthetic objects

---
P. Pinggera et al., Lost and Found: Detecting Small Road Hazards for Self-Driving Vehicles, IROS 2016
R. Chan et al., SegmentMeIfYouCan: A Benchmark for Anomaly Segmentation, NeurIPS Datasets and Benchmarks 2023
T.H. VU et al., The BRAVO Semantic Segmentation Challenge Results in UNCV2024

Case 2 - Data scarcity

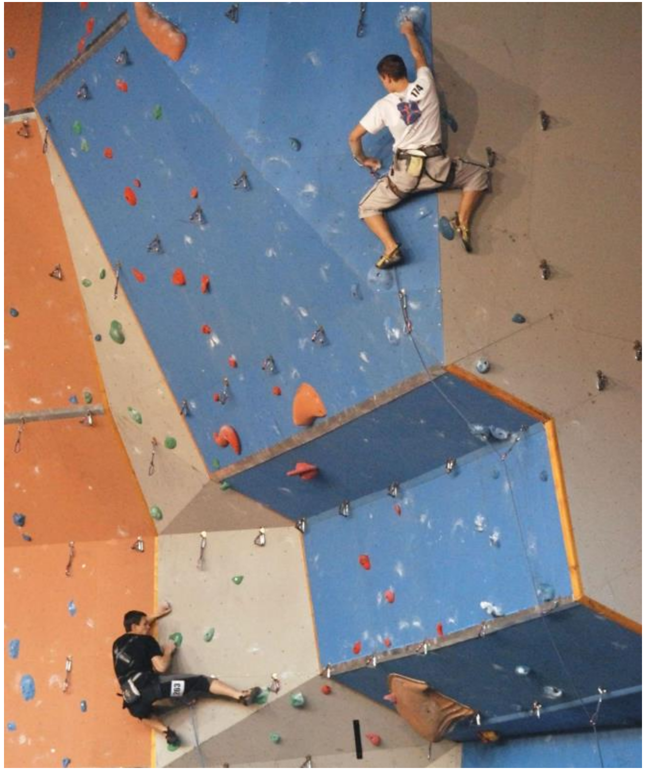


Also causing knowledge / epistemic uncertainty

A. Malinin, Uncertainty Estimation in Deep Learning with application to Spoken Language Assessment, PhD Thesis 2019

Case 2 - Data scarcity

Train samples



Test samples: unseen variations of known classes

S. Kreiss et al., PifPaf: Composite Fields for Human Pose Estimation, CVPR 2019

Knowledge uncertainty




- Knowledge uncertainty is caused by the lack of knowledge about the process that generated the data.

- It can be reduced with additional and sufficient training data (reducible uncertainty*)

*reducible, but not completely

Knowing which source of uncertainty predominates can be useful for:

- active learning, reinforcement learning (knowledge uncertainty)
- new data acquisition (knowledge uncertainty)
- distribution shifts (knowledge uncertainty)

- decide to fall-back to a complementary sensor, human, etc. (data uncertainty)
- ambiguity or multiple predictions (data uncertainty)
- failure detection (predictive uncertainty)

The described data and knowledge uncertainty sources are idealized:

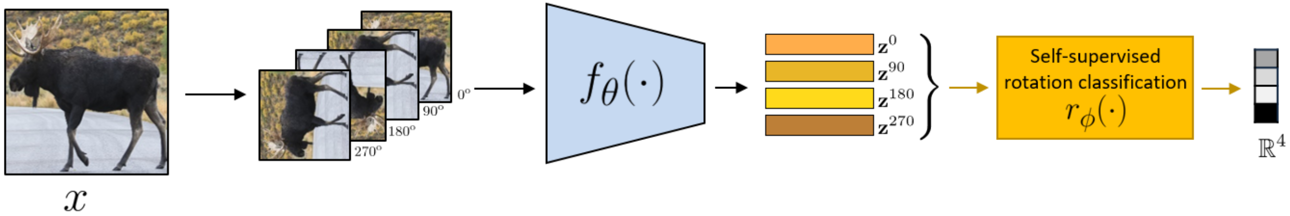
- In practice, real data have both uncertainties intermingled and accumulating in predictive/total uncertainty.
- Most models do not always satisfy conditions for data uncertainty estimation, e.g., overconfidence

Foundation models:
train once, use many times

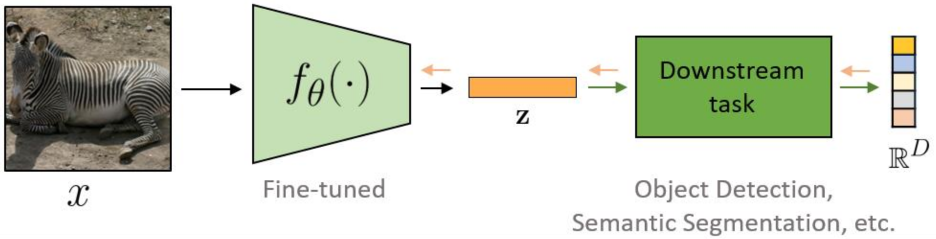
“A foundation model is any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks.” (Bommasani et al., 2021)

Self-supervised learning pipelines in the 2010s

Stage 1: Pretrain network on pretext task (without human labels)



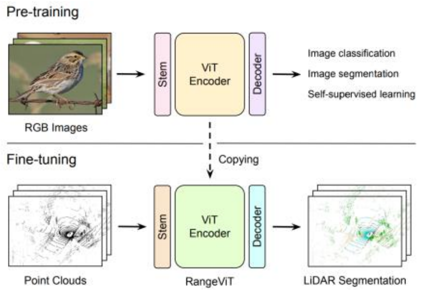
Stage 2: Fine-tune network for new task with fewer labels



New use-cases possible with recent foundation models

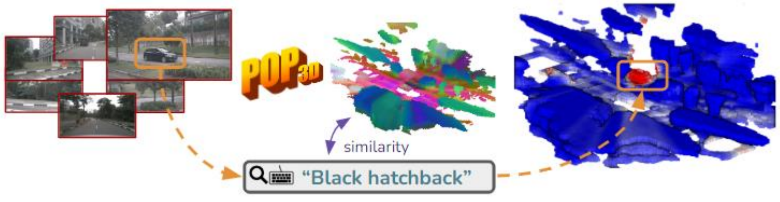
Stage 2: Can also be distillation, data mining, active learning, model initialization …

reuse pretrained backbones on other modalities



RangeViT [CVPR’23]

text-driven 3D retrieval from cameras

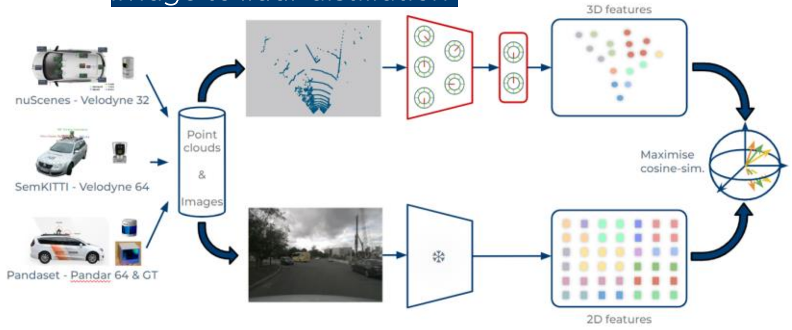


POP-3D [NeurIPS’23]

unsupervised semantic segmentation

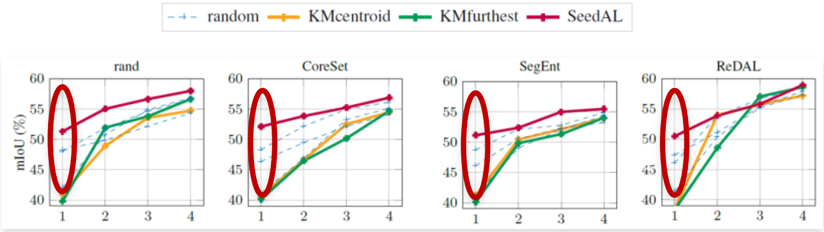
Drive&Segment [ECCV’22]

image to lidar distillation



ScaLR [CVPR’24]

kickstart active learning



SeedAL [ICCV’23]

Make Me a BNN: A Simple Strategy for Estimating Bayesian Uncertainty from Pre-trained Models

Gianni Franchi, Olivier Laurent, Maxence Leguéry,
Andrei Bursuc, Andrea Pilzer, Angela Yao

CVPR 2024

Notations

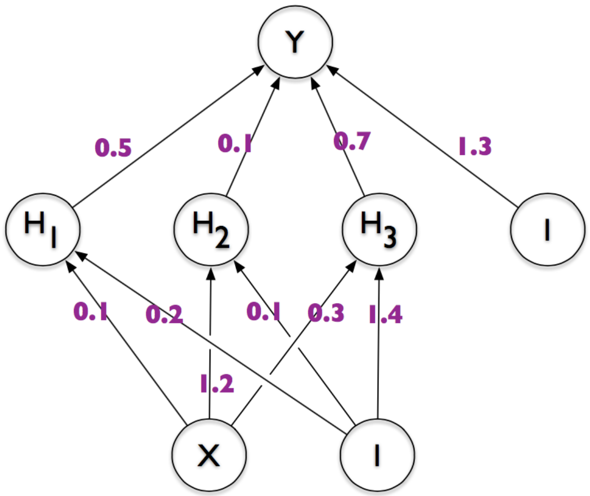
- We consider a training dataset $\mathcal{D} = \{(x_i, y_i)\}$ with N samples and labels
- We view our network f as a probabilistic model with $f_\omega(x_i) = P(y_i \mid x_i, \omega)$
- The model posterior $p(\omega \mid \mathcal{D})$ captures the uncertainty in ω and we compute it during training:

$$\overbrace{p(\omega \mid \mathcal{D})}^{\text{posterior}} = \frac{\overbrace{p(\mathcal{D} \mid \omega)}^{\text{likelihood}}\overbrace{p(\omega)}^{\text{prior}}}{p(\mathcal{D})}$$

- most models find a single set or parameters to maximize the probability on conditioned data

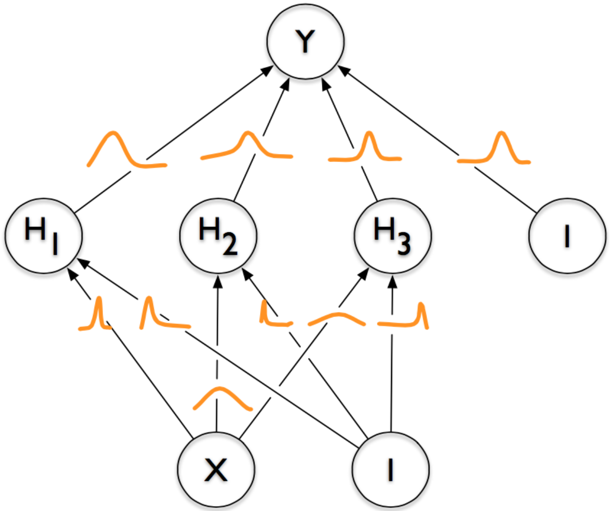
$$\omega^* = \arg\max_\omega p(\omega \mid x, y) \approx \arg\max_\omega \sum_{x,y \in \mathcal{D}} \log p(y \mid x, \omega) + \log p(\omega)$$

Standard Neural Network



- Parameters represented by single, fixed values (point estimates)
- Conventional approaches to training NNs can be interpreted as approximations to the full Bayesian method (equivalent to MLE or MAP estimation)

Bayesian Neural Network



- Parameters represented by distributions
- For Gaussian priors: each parameter consists of a pa (μ, σ) describing a distribution over it (2x more parameters)

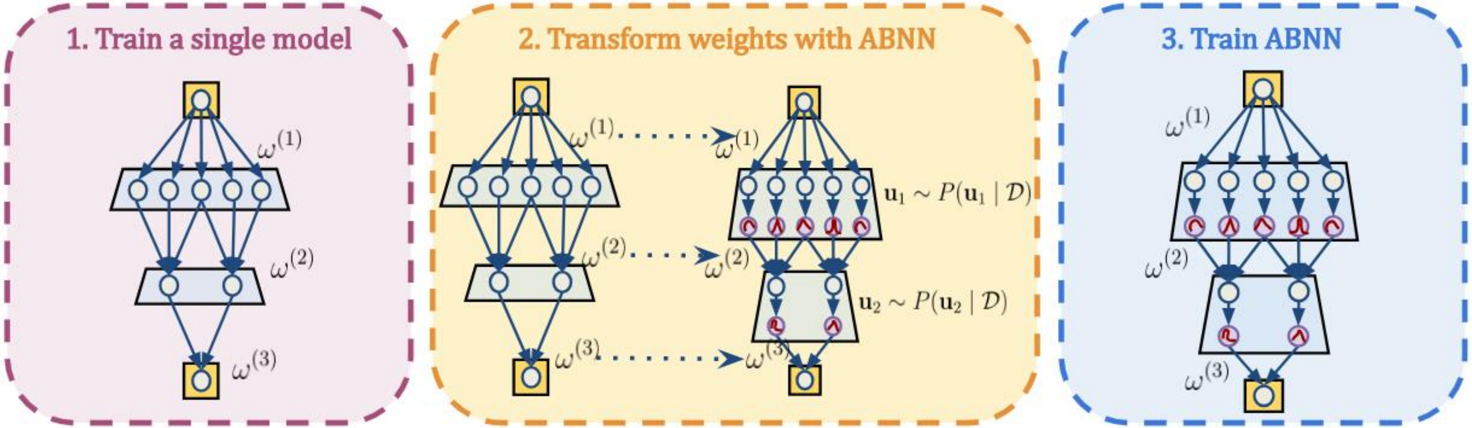
C. Blundell et al., Weight Uncertainty in Neural Networks, ICML 2015

- Bayesian Neural Networks (BNNs) are easy to formulate, but difficult to perform inference in.

- Modern BNNs are trained with variational inference (reparameterization trick), but unstable to train at scale (even for CNNs)

- Ensemble approaches have been historically popular but prohibitive in foundation models

Turning a DNN into a BNN



- From a pretrained network we derive a BNN by adjusting normalization layers to become stochastic -> Bayesian Normalization Layer (BNL)
- The final step involved finetuning the Adaptable BNN over a limited number of steps

Turning a DNN into a BNN

- Formally our BNN relies on a new layer BNL (j-th layer:

$$\mathbf{u}_j = \mathbf{BNL}\left(W^{(j)}\mathbf{h}_{j-1}\right) \qquad \mathbf{a}_j = a(\mathbf{u}_j)$$

$$\mathbf{BNL}(\mathbf{h}_j) = \frac{\mathbf{h}_j - \hat{\mu}_j}{\hat{\sigma}_j} \times \gamma_j(1 + \epsilon_j) + \beta_j \qquad \epsilon_j \sim \mathcal{N}(\mathbf{0}, \mathbb{1})$$

γ_j , β_j are learnable parameters

- This can be seen as adding a Gaussian dropout on normalization layer and finetuning the DNN

$$\mathcal{L}(\omega) = \mathcal{L}_{\mathrm{MAP}}(\omega) + \mathcal{E}(\omega)$$

$\mathcal{E}(\omega)$ class-dependent perturbation

- We train multiple ABNN to cover multiple posterior modes and sample + ensemble at runtime

Classification results

		CIFAR-10					CIFAR-100					
	Method	Acc ↑	NLL ↓	AUPR ↑	AUC ↑	FPR95 ↓	Acc ↑	NLL ↓	AUPR ↑	AUC ↑	FPR95 ↓	Time (h) ↓
ResNet-50	Single Model	95.1	0.211	95.2	91.9	23.6	78.3	0.905	87.4	77.9	57.6	**1.7**
	BatchEnsemble	93.9	0.255	94.7	91.3	20.1	66.6	1.788	85.2	74.6	60.6	17.2
	LPBNN	94.3	0.231	92.7	86.7	54.9	78.5	1.02	88.2	77.8	73.5	17.2
	MCDropout	94.4	0.190	93.1	86.9	43.8	76.9	0.858	87.8	77.1	64.1	**1.7**
	MCBN	95.0	0.168	95.7	92.6	20.1	78.4	0.83	86.8	77.5	57.7	**1.7**
	Deep Ensembles	**96.0**	**0.136**	**97.0**	**94.7**	**80.9**	**0.713**	**2.6**	89.2	80.8	52.5	6.8
	Laplace	95.3	0.160	96.0	93.3	78.2	0.99	14.2	89.2	81.0	51.8	**1.7**
	ABNN	95.0	0.160	96.5	93.9	17.5	77.8	0.828	**90.0**	**82.0**	**51.3**	2.0
WideResNet-28×10	Single Model	95.4	0.200	96.1	93.2	20.4	80.3	0.963	81.0	64.2	80.1	**4.2**
	BatchEnsemble	95.6	0.206	95.5	92.5	22.1	82.3	0.835	88.1	78.2	69.8	25.6
	LPBNN	95.1	0.249	95.4	91.2	29.5	79.7	0.831	79.0	70.1	71.4	23.3
	MCDropout	95.7	0.138	96.2	93.5	12.8	79.2	0.758	**89.4**	**80.1**	58.6	**4.2**
	MCBN	95.5	**0.133**	96.5	94.2	14.6	80.4	0.749	80.4	67.8	63.1	**4.2**
	Deep Ensembles	**95.8**	0.143	**97.8**	**96.0**	**82.5**	0.903	22.9	81.6	67.9	71.3	16.6
	Laplace	95.6	0.151	95.0	90.7	31.9	80.1	0.942	83.4	72.1	59.9	**4.2**
	ABNN	94.5	0.171	0.7	96.8	94.6	80.0	**0.734**	86.7	75.7	59.4	5.0

- ABNN improves uncertainty quantification with small computational overhead
- Most of the gains are linked to improved knowledge uncertainty (OOD detection)

Semantic segmentation results

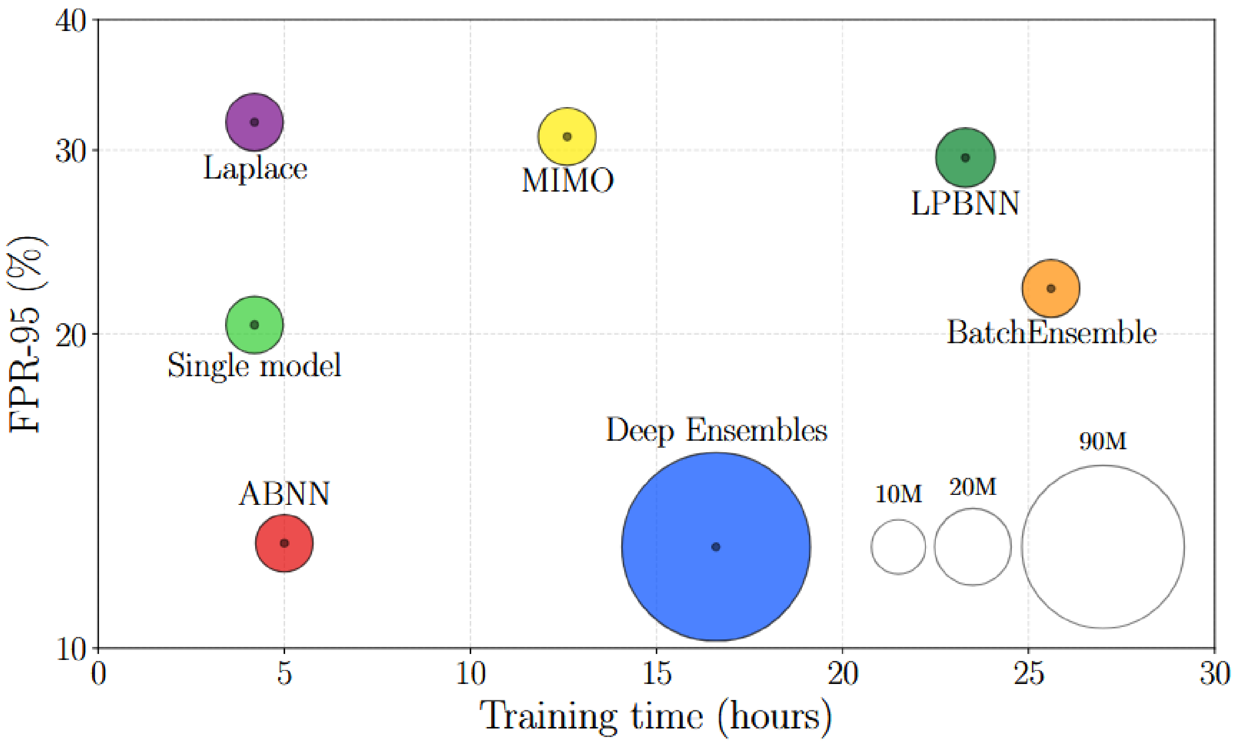
Dataset	Method	mIoU ↑	AUPR ↑	AUC ↑	FPR95 ↓	ECE ↓
StreetHazards	Single Model	53.90	6.91	86.60	35.74	6.52
	TRADI	52.46	6.93	87.39	38.26	6.33
	Deep Ensembles	55.59	**8.32**	87.94	**30.29**	5.33
	MIMO	55.44	6.90	87.38	32.66	5.57
	BatchEnsemble	**56.16**	7.59	88.17	32.85	6.09
	LP-BNN	54.50	7.18	88.33	32.61	**5.20**
	ABNN (ours)	53.82	7.85	**88.39**	32.02	6.09
BDD-Anomaly	Single Model	47.63	4.50	85.15	28.78	17.68
	TRADI	44.26	4.54	84.80	36.87	16.61
	Deep Ensembles	**51.07**	5.24	84.80	**28.55**	14.19
	MIMO	47.20	4.32	84.38	35.24	16.33
	BatchEnsemble	48.09	4.49	84.27	30.17	16.90
	LP-BNN	49.01	4.52	85.32	29.47	17.16
	ABNN (ours)	48.76	**5.98**	**85.74**	29.01	**14.03**
MUAD	Single Model	57.32	26.04	86.24	39.43	6.07
	MC-Dropout	55.62	22.25	84.39	45.75	6.45
	Deep Ensembles	58.29	**28.02**	87.10	37.60	5.88
	BatchEnsemble	57.10	25.70	86.90	38.81	6.01
	MIMO	57.10	24.18	86.62	34.80	5.81
	ABNN (ours)	**61.96**	24.37	**91.55**	**21.68**	**5.58**

- Different distribution shifts (adverse weather, unknown objects)
- ABNN also performs well here

Larger encoders

	Method	Acc ↑	ECE ↓	AUPR ↑	AUC ↑	FPR95 ↓
ResNet-50	Single Model	77.8	12.1	18.0	80.9	68.6
	BatchEnsemble	75.9	**3.5**	**20.2**	81.6	66.5
	MIMO ($\rho = 1$)	77.6	14.7	18.4	81.6	66.8
	Deep Ensembles	79.2	23.3	19.6	**83.4**	**62.1**
	Laplace	**80.4**	44.3	13.9	75.9	82.8
	ABNN	79.5	9.65	17.8	82.0	65.2
ViT	Single Model	80.0	5.2	19.5	84.1	58.5
	Deep Ensembles	**81.7**	13.5	21.7	**85.5**	60.3
	Laplace	81.0	10.8	**22.1**	83.1	70.6
	ABNN	80.6	**4.32**	21.7	85.4	**55.1**

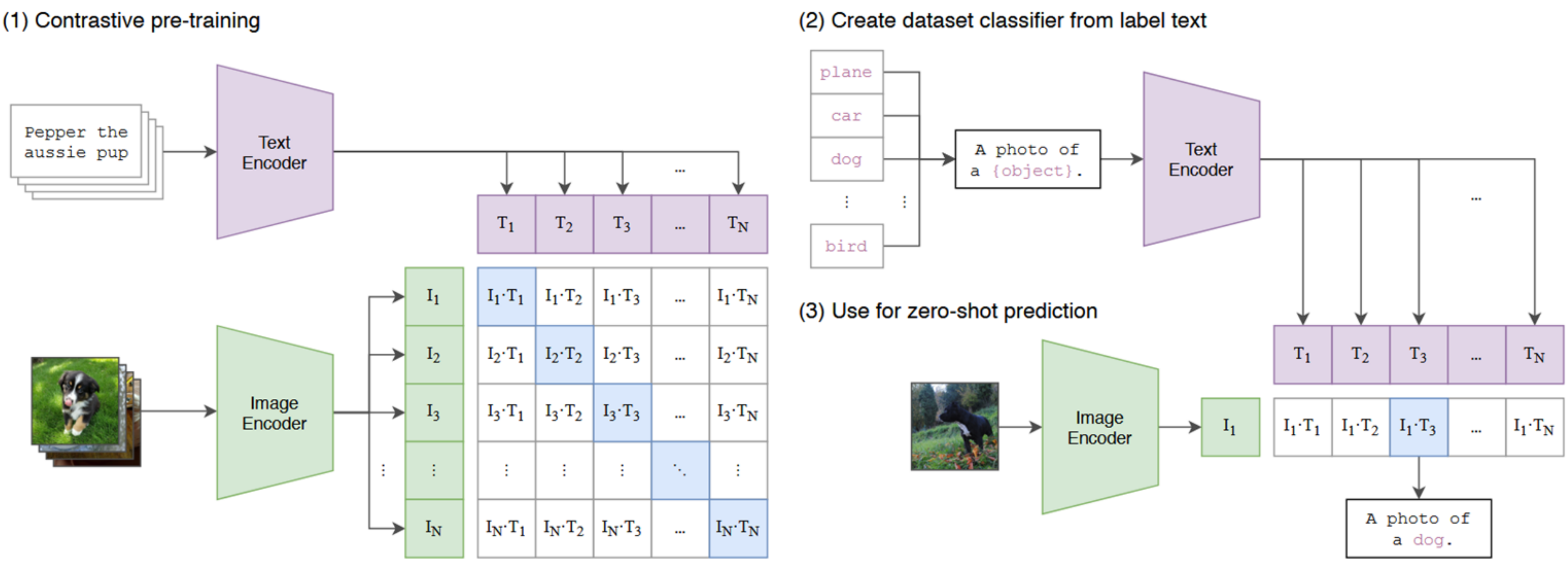
- Improving upon ViT pretrained on ImageNet-21K



Code available at: https://github.com/ENSTA-U2IS-AI/torch-uncertainty

Vision-Language Models (VLMs)

CLIP

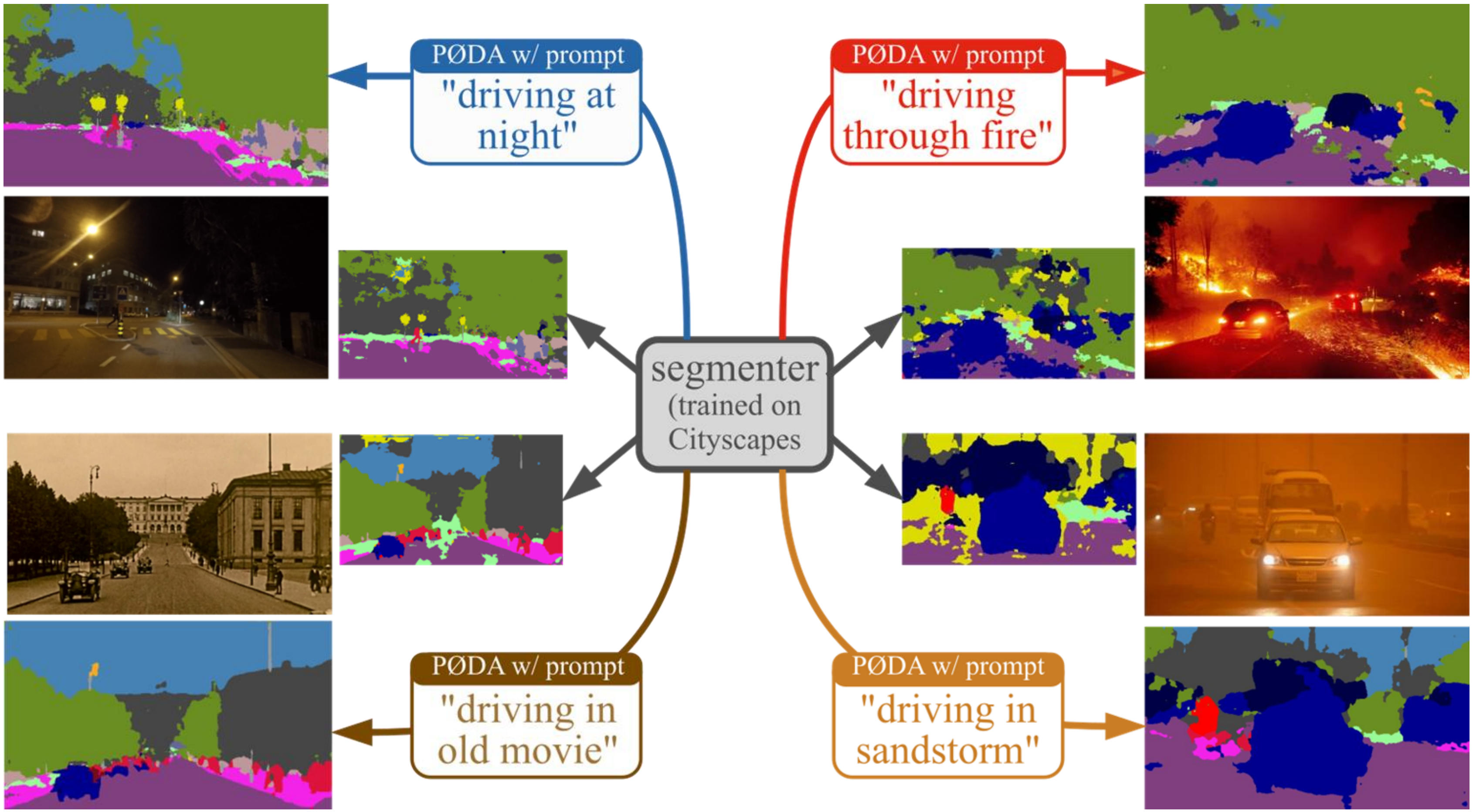


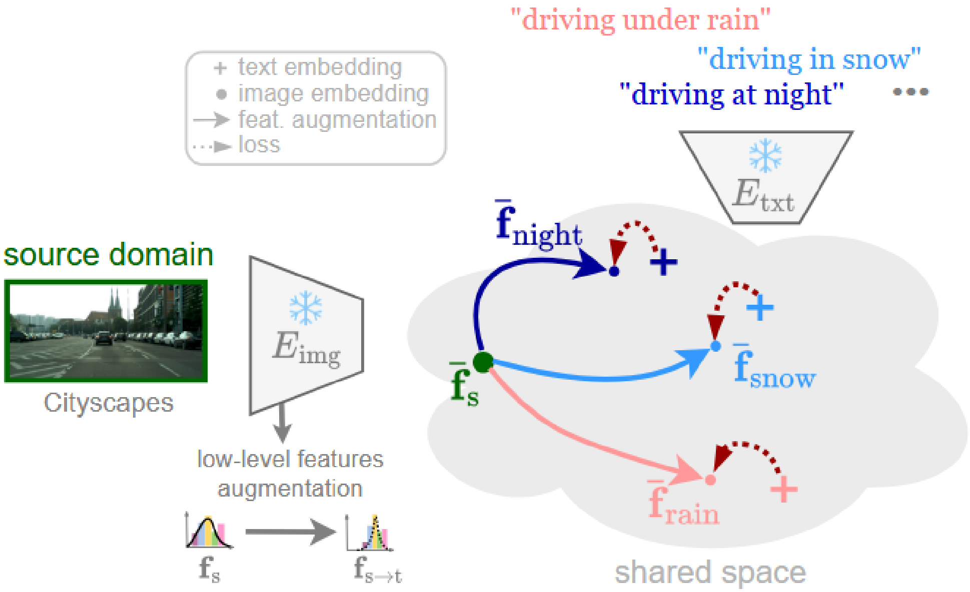
A. Radford, Learning Transferable Visual Models From Natural Language Supervision, ICML 2021

PØDA: Prompt-driven Zero-shot Domain Adaptation

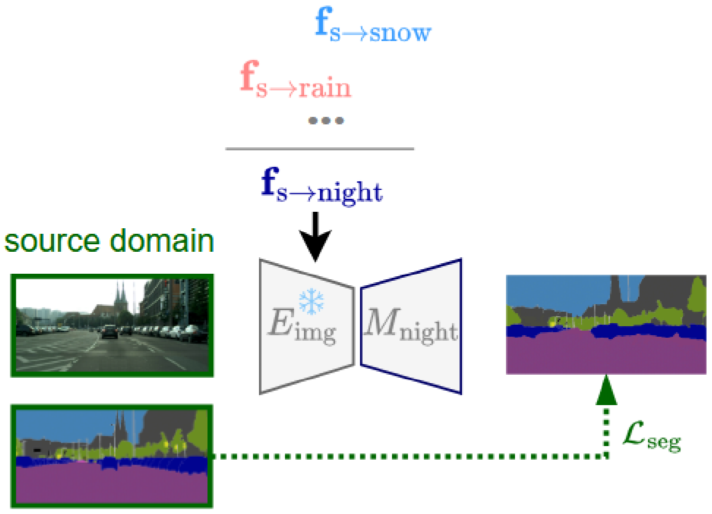
Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc,
Patrick Pérez, Raoul de Charette

IJCV 2023, journal review


PØDA w/ prompt
"driving at night"
PØDA w/ prompt
"driving through fire"
segmenter (trained on Cityscapes
PØDA w/ prompt
"driving in old movie"
PØDA w/ prompt
"driving in sandstorm"

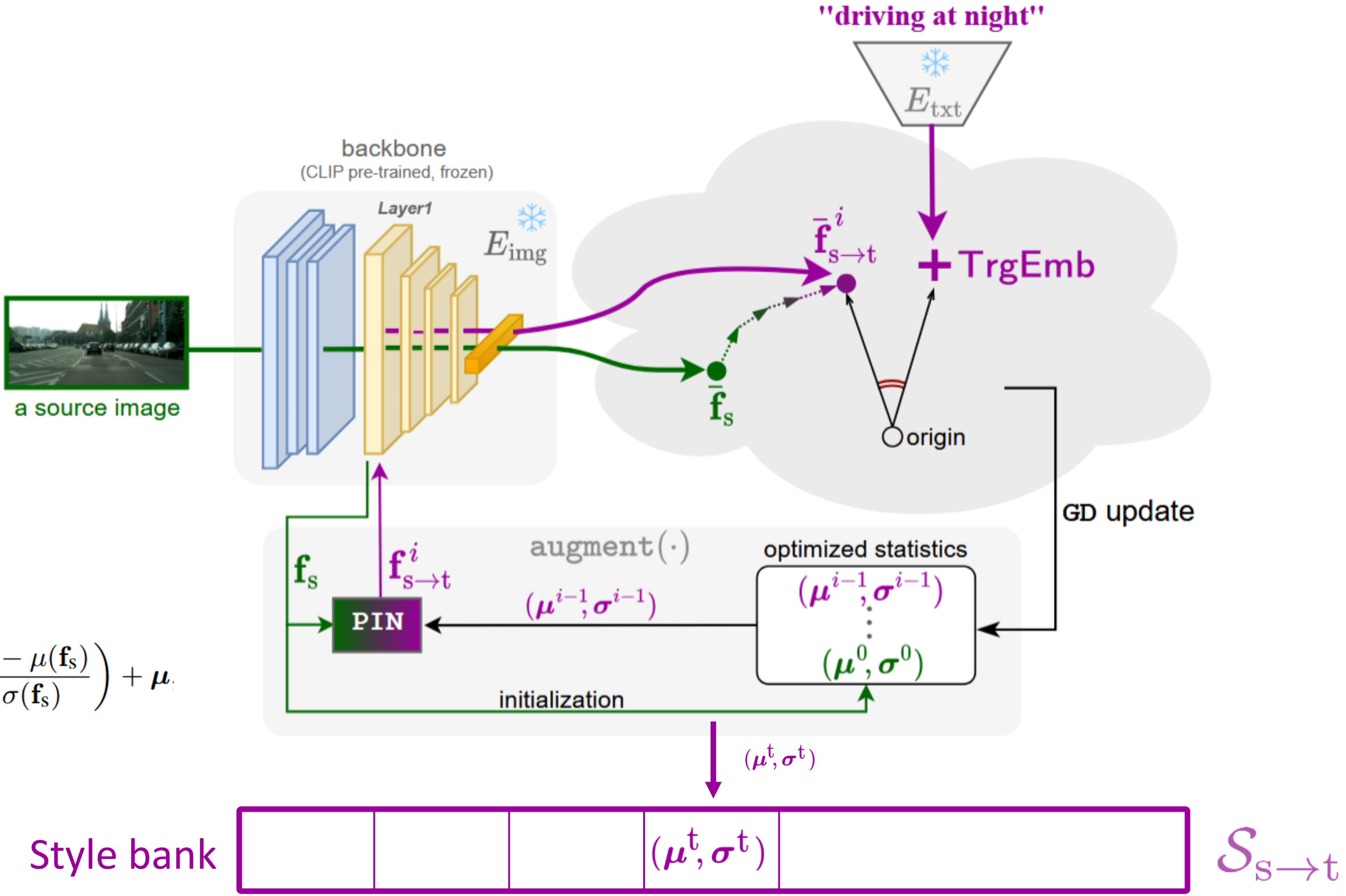


Prompt-driven feature augmentation



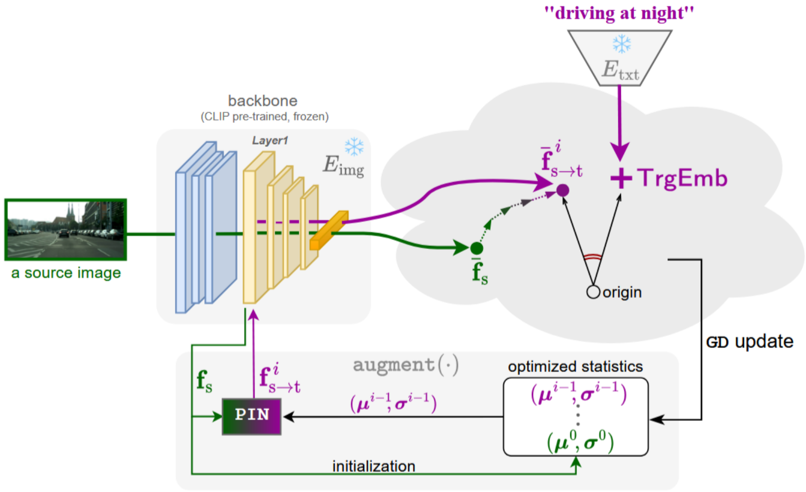
Zero-shot Domain Adaptation

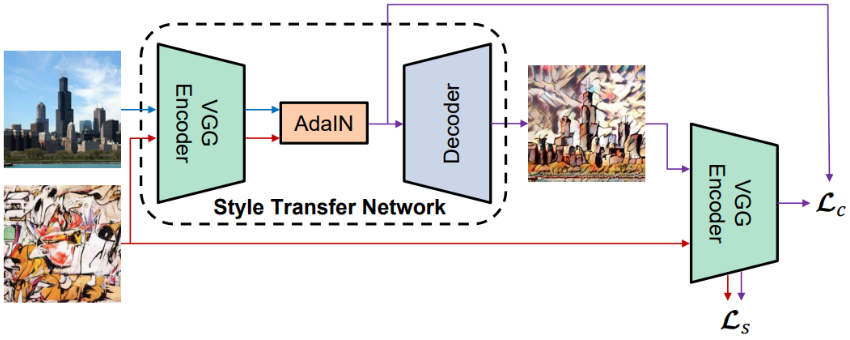
Prompt-Driven Instance Normalization (PIN)



$$\mathtt{PIN}(\mathbf{f}_{\text{s}}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \boldsymbol{\sigma}\left(\frac{\mathbf{f}_{\text{s}} - \mu(\mathbf{f}_{\text{s}})}{\sigma(\mathbf{f}_{\text{s}})}\right) + \boldsymbol{\mu}$$

PIN vs AdaIN





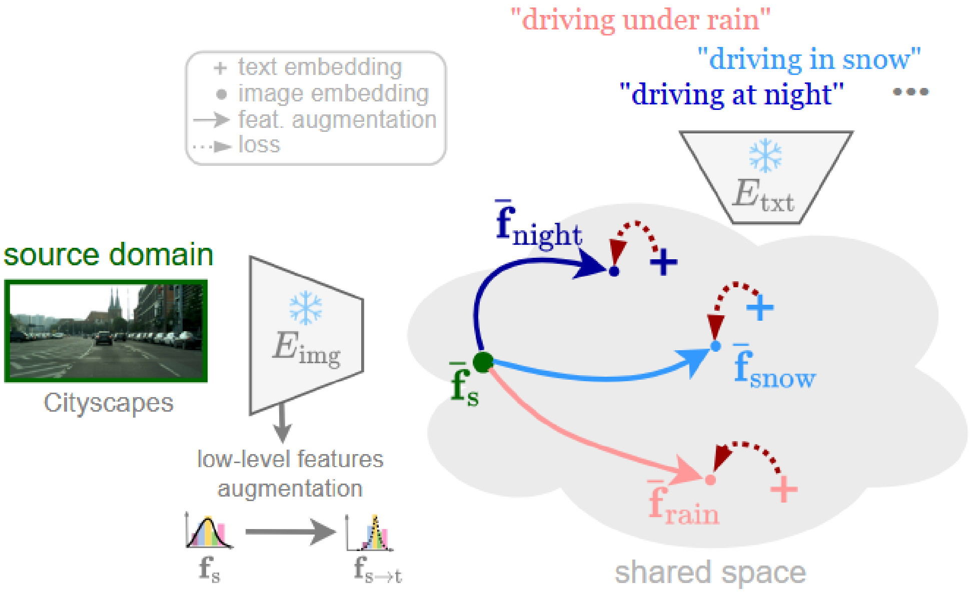
$$\mathtt{PIN}(\mathbf{f}_s, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \boldsymbol{\sigma}\left(\frac{\mathbf{f}_s - \mu(\mathbf{f}_s)}{\sigma(\mathbf{f}_s)}\right) + \boldsymbol{\mu},$$

$$\mathtt{AdaIN}(\mathbf{f}_s, \mathbf{f}_t) = \sigma(\mathbf{f}_t)\left(\frac{\mathbf{f}_s - \mu(\mathbf{f}_s)}{\sigma(\mathbf{f}_s)}\right) + \mu(\mathbf{f}_t)$$

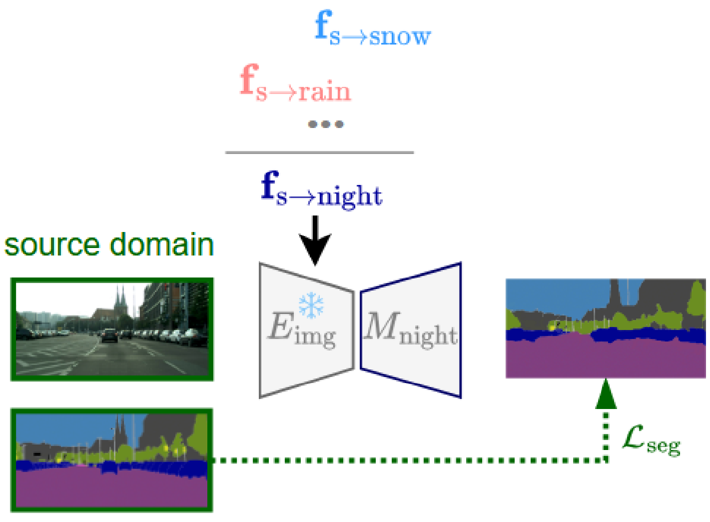
- Very similar formalism
- AdaIN uses target image, PIN uses features via target prompt

---

X. Huang & S. Belongie, Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization, ICCV 2017



Prompt-driven feature augmentation



Zero-shot Domain Adaptation

15 min

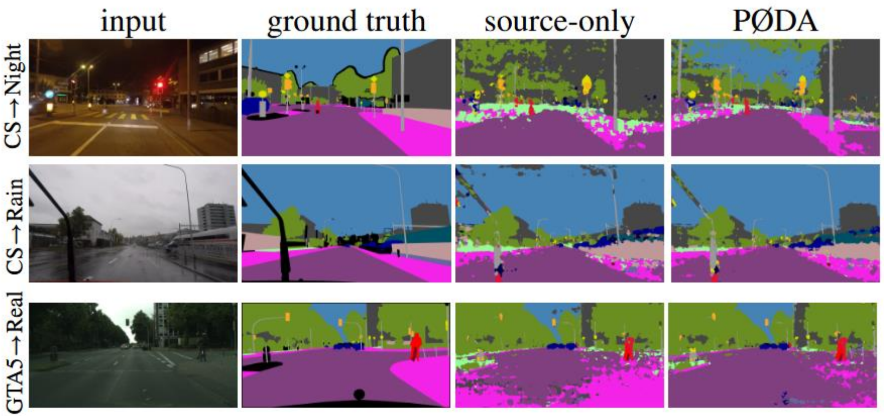
Training on Cityscapes

10 min

PØDA:
Prompt-driven Zero-shot Domain Adaptation

(no audio)

Results



Proxies

Source	Target eval.	Method	mIoU[%]
	TrgPrompt = "driving at night"		
CS	ACDC Night	source-only	18.31
		CLIPstyler	21.38 ±0.36
		PØDA	**25.03** ±0.48
	TrgPrompt = "driving in snow"		
	ACDC Snow	source-only	39.28
		CLIPstyler	41.09 ±0.17
		PØDA	**43.90** ±0.53
	TrgPrompt = "driving under rain"		
	ACDC Rain	source-only	38.20
		CLIPstyler	37.17 ±0.10
		PØDA	**42.31** ±0.55
	TrgPrompt = "driving in a game"		
	GTA5	source-only	39.59
		CLIPstyler	38.73 ±0.16
		PØDA	**41.07** ±0.48
	TrgPrompt = "driving"		
GTA5	CS	source-only	36.38
		CLIPstyler	31.50 ±0.21
		PØDA	**40.08** ±0.52

+1% to +7%

Prompt design

give me 5 prompts that have the same exact meaning as "{prompt}"

Chat GPT

give me 5 random prompts of length from 3 to 6 words describing a random photo

Method	ACDC Night	ACDC Snow	ACDC Rain	GTA5
Source only	18.31	39.28	38.20	39.59
Trg	"driving at night"	"driving in snow"	"driving under rain"	"driving in a game"
	25.03 ±0.48	43.90 ±0.53	42.31 ±0.55	41.07 ±0.48
ChatGPT-generated, Relevant →	"operating a vehicle after sunset"	"operating a vehicle in snowy conditions"	"operating a vehicle in wet conditions"	"piloting a vehicle in a virtual world"
	24.38 ±0.37	44.33 ±0.36	42.21 ±0.47	41.25 ±0.40
	"driving during the nighttime hours"	"driving on snow-covered roads"	"driving on rain-soaked roads"	"controlling a car in a digital simulation"
	25.22 ±0.64	43.56 ±0.62	**42.51** ±0.33	41.19 ±0.14
	"navigating the roads in darkness"	"piloting a vehicle in snowy terrain"	"navigating through rainfall while driving"	"maneuvering a vehicle in a computerized racing experience""
	24.73 ±0.47	**44.67** ±0.18	41.11 ±0.69	40.34 ±0.49
	"driving in low-light conditions"	"driving in wintry precipitation"	"driving in inclement weather"	"operating a transport in a video game environment"
	24.68 ±0.34	43.11 ±0.56	40.68 ±0.37	41.34 ±0.42
	"travelling by car after dusk"	"travelling by car in a snowstorm"	"travelling by car during a downpour"	"navigating a machine through a digital driving simulation"
	24.89 ±0.24	43.83 ±0.17	42.05 ±0.35	**41.86** ±0.10
	24.82	*43.90*	*41.81*	*41.18*
ChatGPT-generated, ← Irrelevant	"mesmerizing northern lights display"			
	20.05 ±0.77	40.07 ±0.66	38.43 ±0.82	37.98 ±0.31
	"playful dolphins in the ocean"			
	20.11 ±0.31	39.87 ±0.26	38.56 ±0.58	37.05 ±0.31
	"breathtaking view from mountaintop"			
	20.65 ±0.33	42.08 ±0.28	40.05 ±0.52	40.09 ±0.23
	"cheerful sunflower field in bloom"			
	21.10 ±0.50	39.85 ±0.68	40.09 ±0.41	37.93 ±0.55
	"dramatic cliff overlooking the ocean"			
	20.09 ±0.98	38.20 ±0.54	38.48 ±0.37	37.57 ±0.46
	"majestic eagle in flight over mountains"			
	20.70 ±0.38	39.60 ±0.27	40.38 ±0.86	38.52 ±0.21
	20.45	*39.95*	*39.33*	*38.19*

Always better

Always worse

Generalization to other tasks

Method	Target	CS→ CS Foggy	DWD-Day Clear → Night Clear	Dusk Rainy	Night Rainy	Day Foggy
DA-Faster [8]	✓	32.0	-	-	-	-
ViSGA [42]	✓	43.3	-	-	-	-
NP+ [15]	✗	46.3	-	-	-	-
S-DGOD [55]	✗	-	36.6	28.2	16.6	33.5
CLIP The Gap [49]	✗	-	36.9	32.3	18.7	38.5
PØDA	✗	**47.3**	**43.4**	**40.2**	**20.5**	**44.4**

Object Detection

"Painting of a bird"

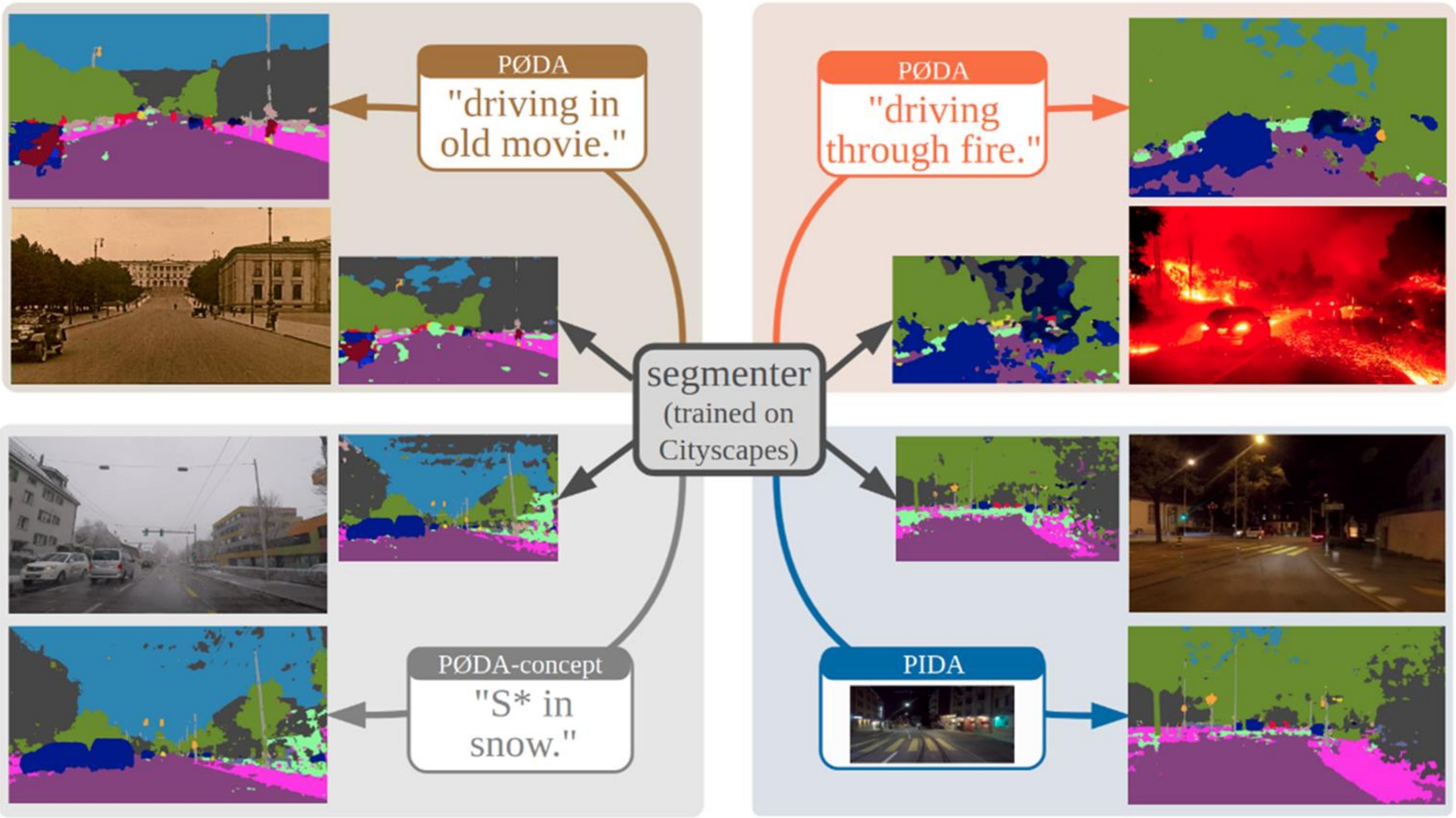
"Blue/Red digits"

Method	CUB-200 paintings	Colored MNIST
src-only	28.90	55.83
PØDA	**30.91** ±0.69	**64.16** ±0.41

Classification

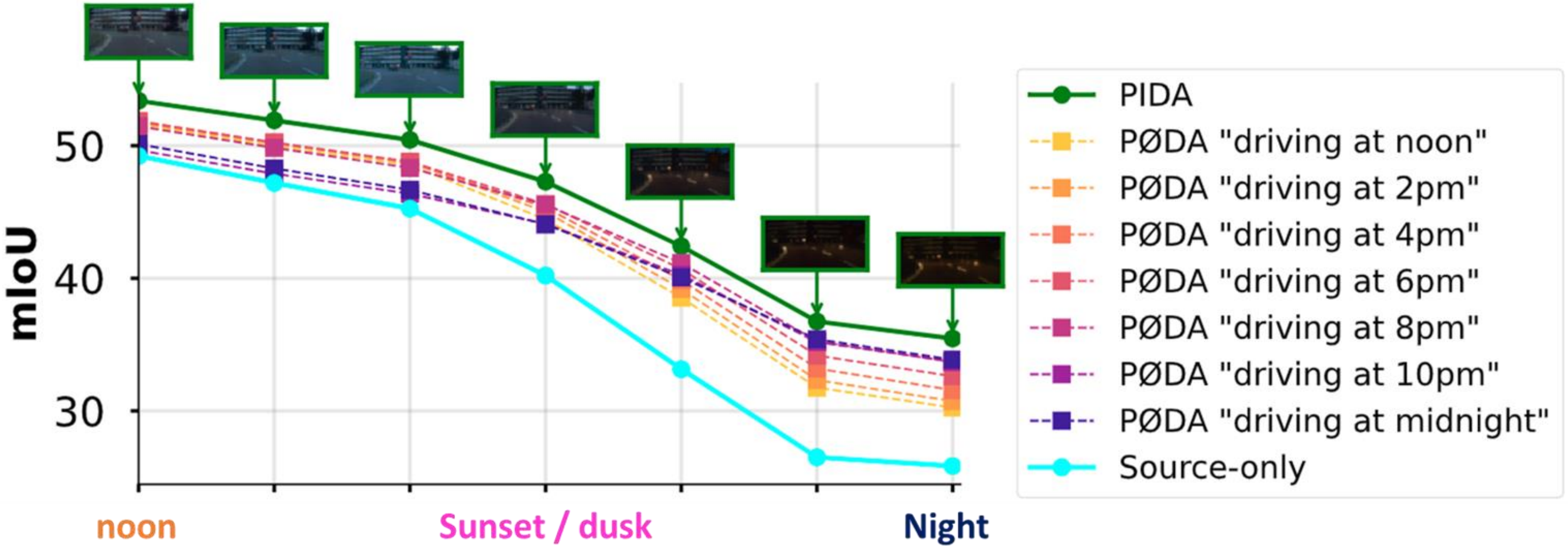
+1% to 8%

+2% to 9%



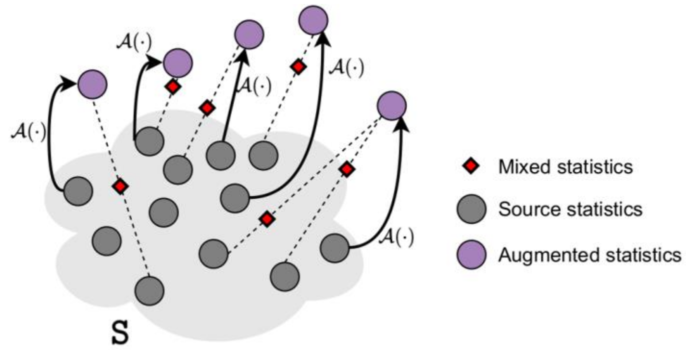
M. Fahes, Domain Adaptation with a Single Vision-Language Embedding, arXiv 2025

Performance at different time-of-day



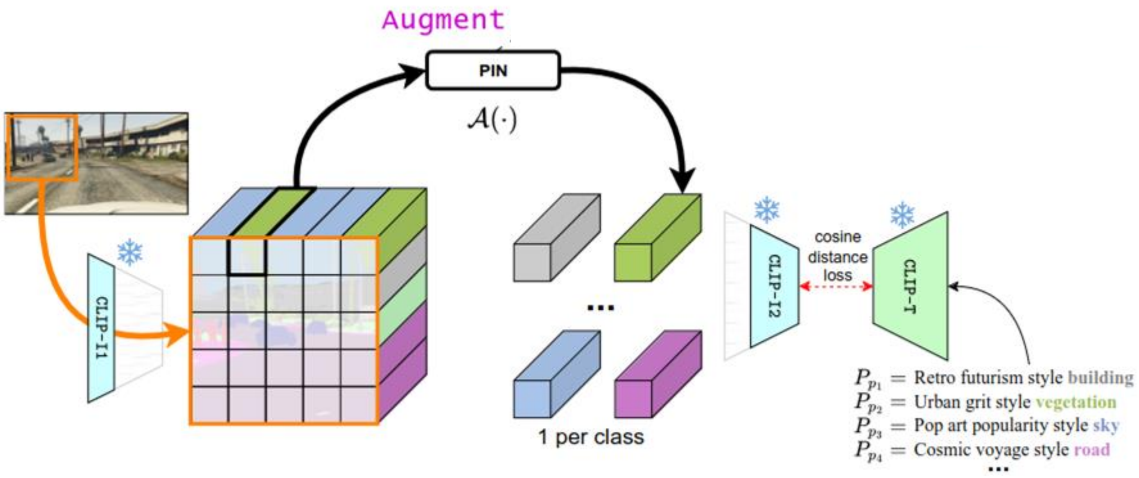
M. Fahes, Domain Adaptation with a Single Vision-Language Embedding, arXiv 2025

Can language boost generalization ?

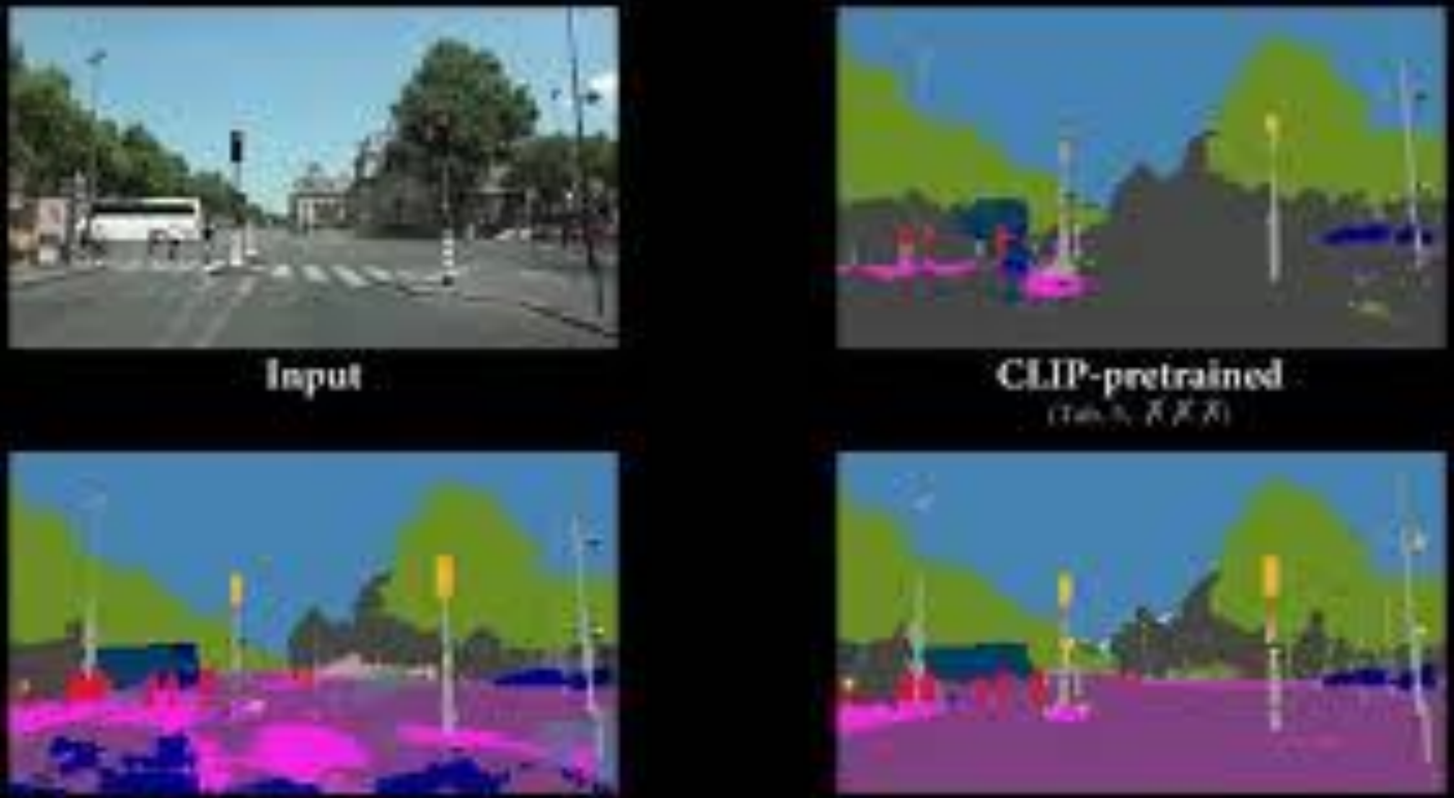


ChatGPT

Retro futurism style building
Urban grit style vegetation
Pop art popularity style sky
Cosmic voyage style road
...



M. Fahes, A Simple Recipe for Language-guided Domain Generalized Segmentation, CVPR 2024

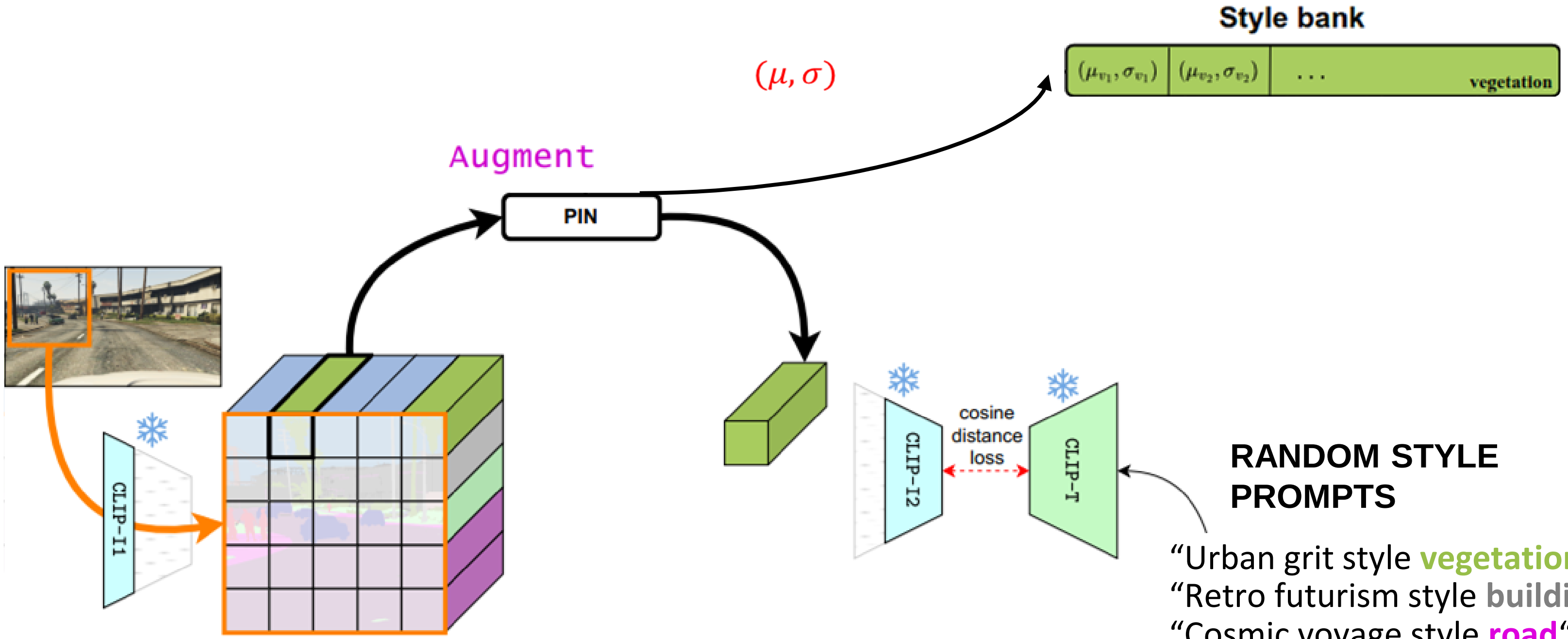

Input
CLIP-pretrained
SHADE
FAMix
Paris

FAMix: A Simple Recipe for Language-guided Domain Generalized Segmentation

Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc,
Patrick Pérez, Raoul de Charette

CVPR 2024

Augment

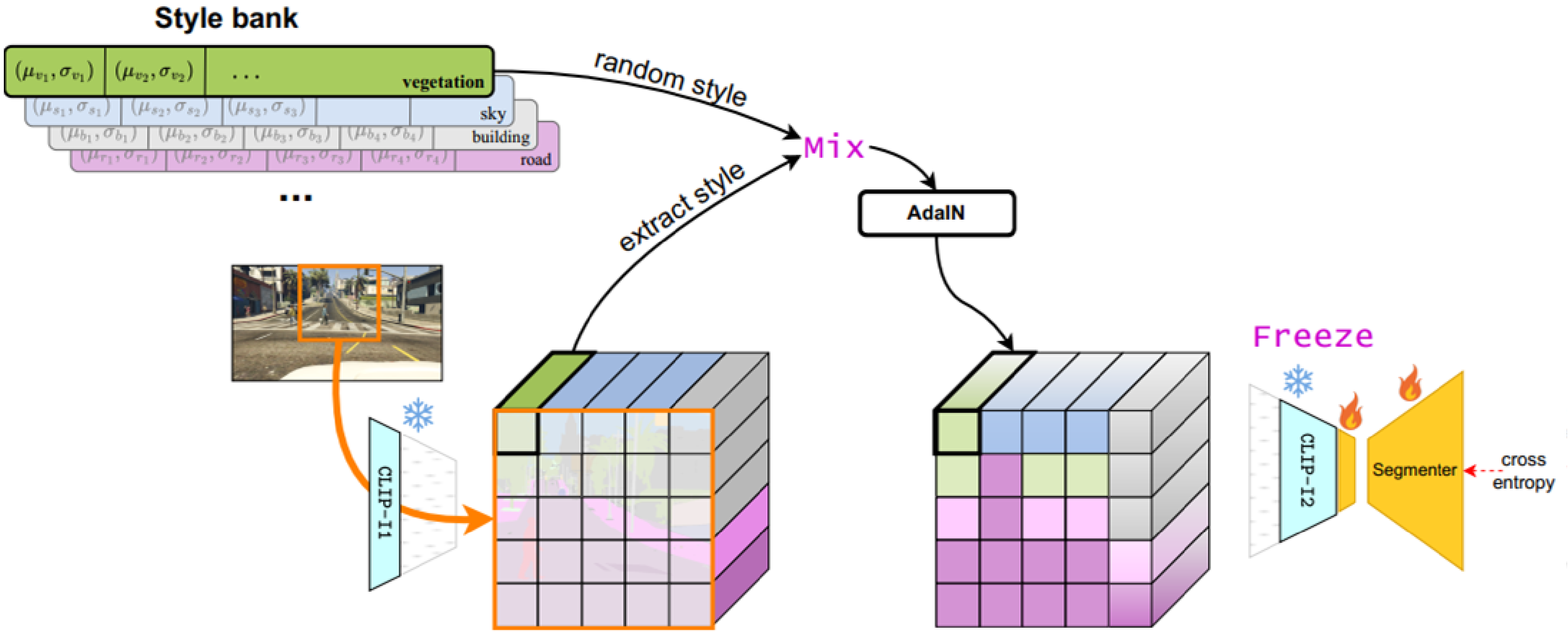


RANDOM STYLE PROMPTS

"Urban grit style **vegetation**"
"Retro futurism style **building**"
"Cosmic voyage style **road**"
...

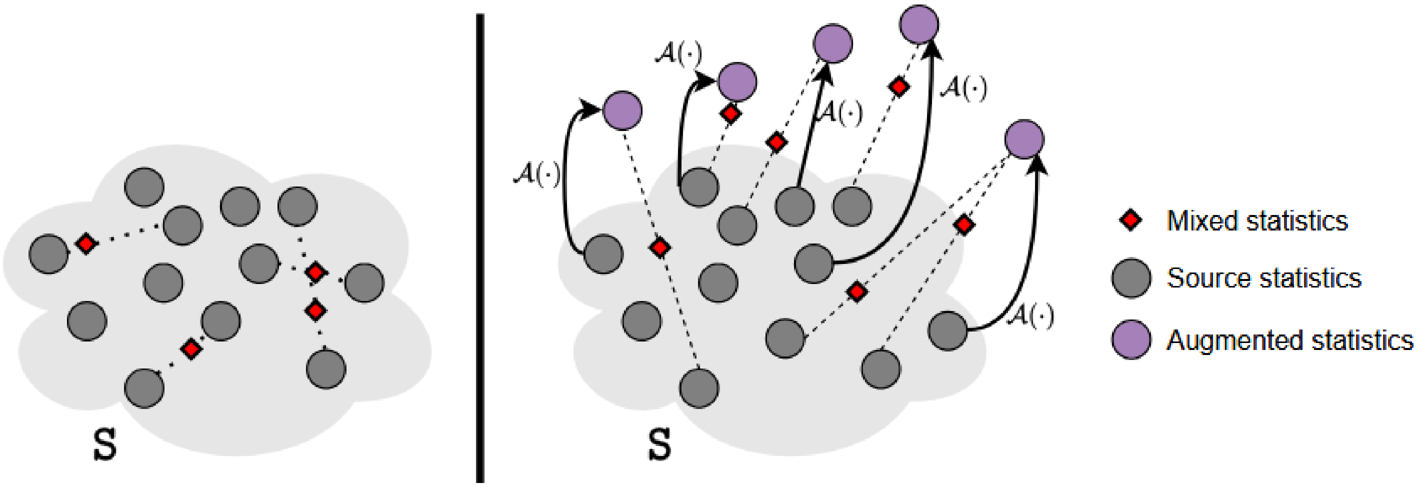
1. Local Style Mining

Augment



2. Training

Augment **Mix**



Patch-wise MixStyle

Ours

+19%

$$\mu_{mix} \leftarrow (1-\boldsymbol{\alpha}) \odot \mu(\mathbf{f}_s^{(ij)}) + \boldsymbol{\alpha} \odot \boldsymbol{\mu}^{(ij)}$$
$$\sigma_{mix} \leftarrow (1-\boldsymbol{\alpha}) \odot \sigma(\mathbf{f}_s^{(ij)}) + \boldsymbol{\alpha} \odot \boldsymbol{\sigma}^{(ij)}$$

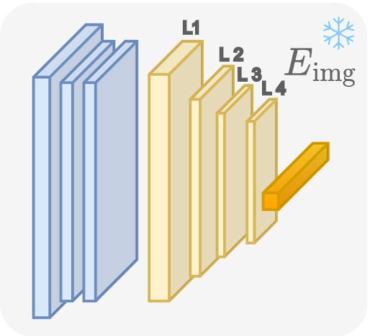
Augment

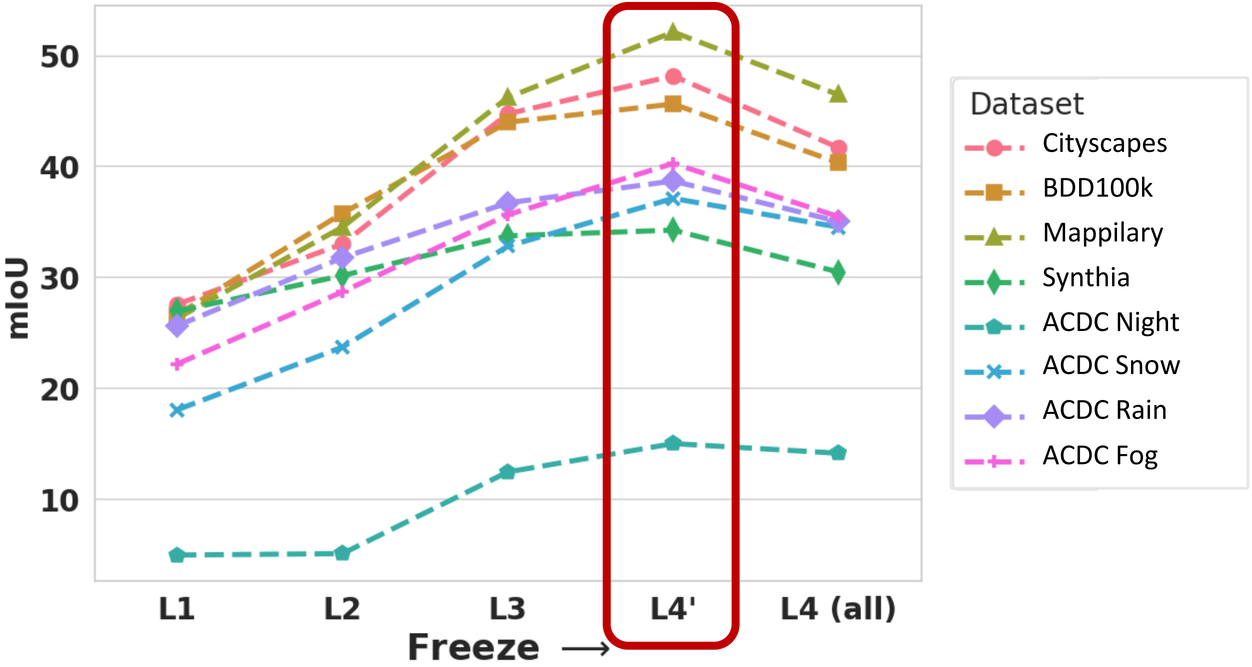
Mix

Freeze

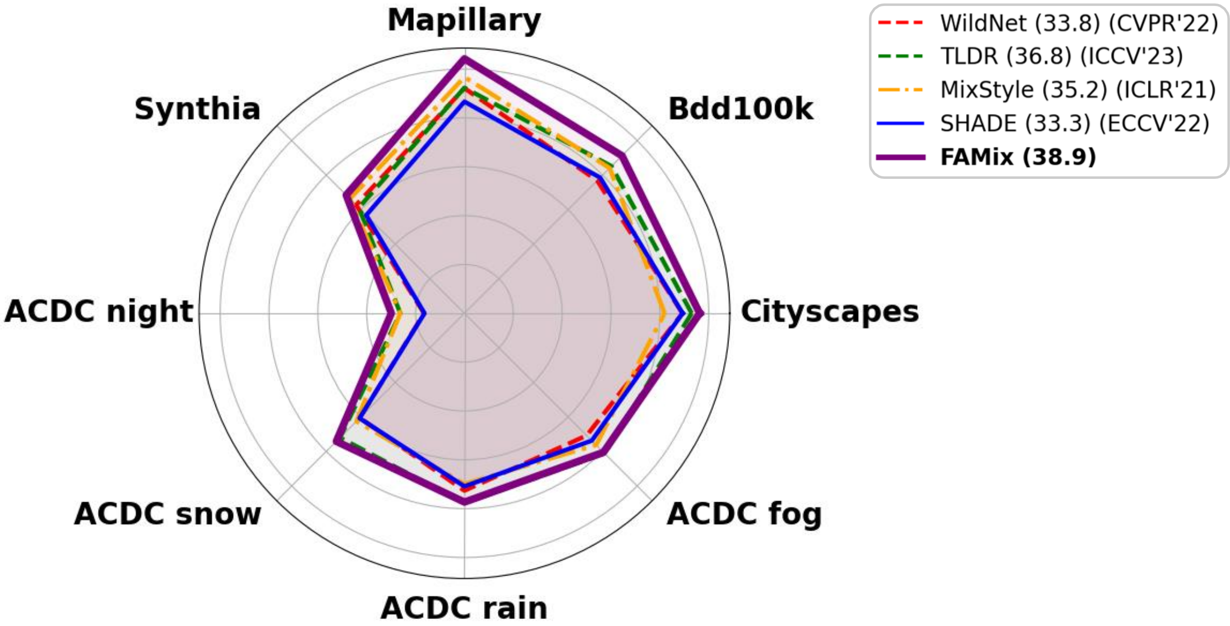
Should we finetune CLIP encoder ?

CLIP image encoder



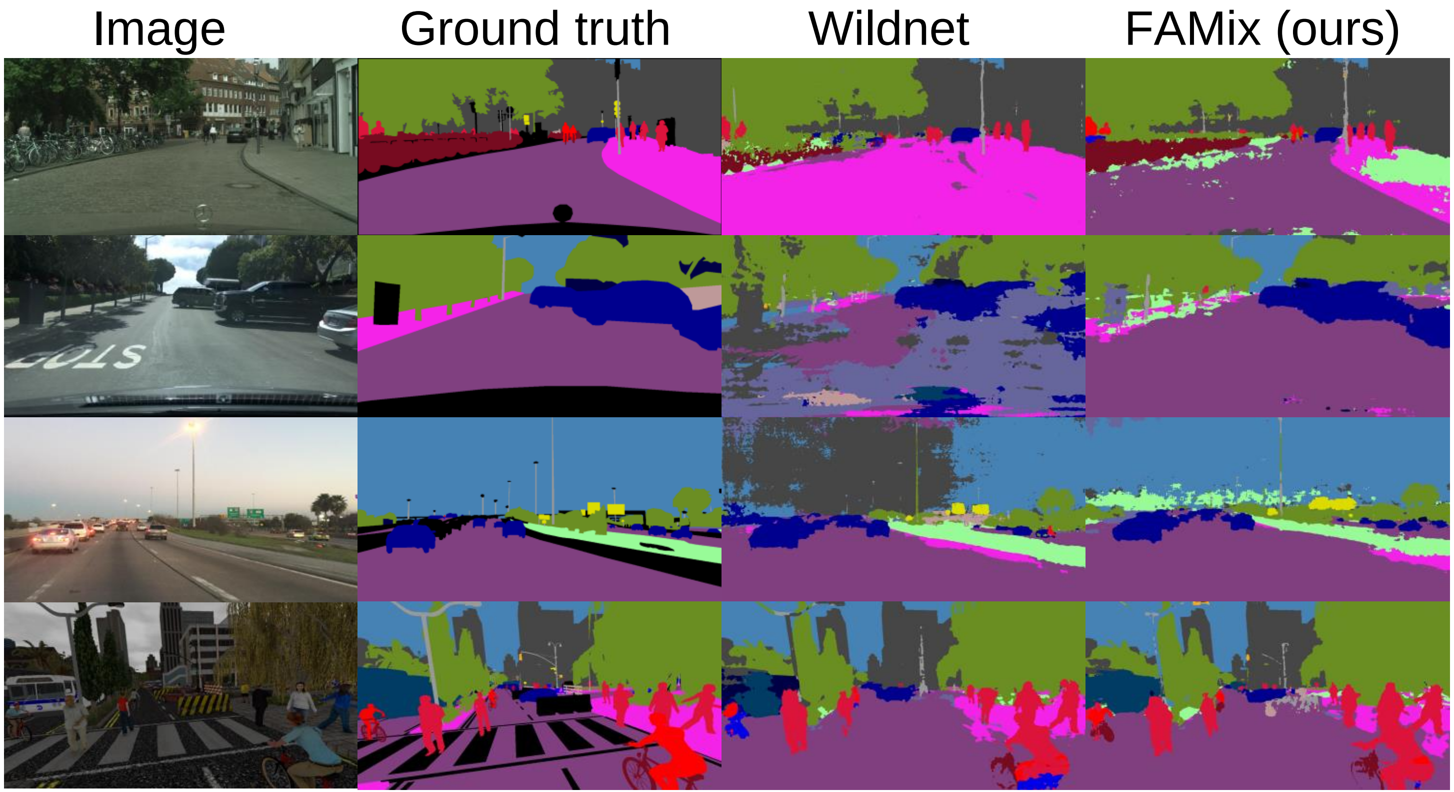


Results

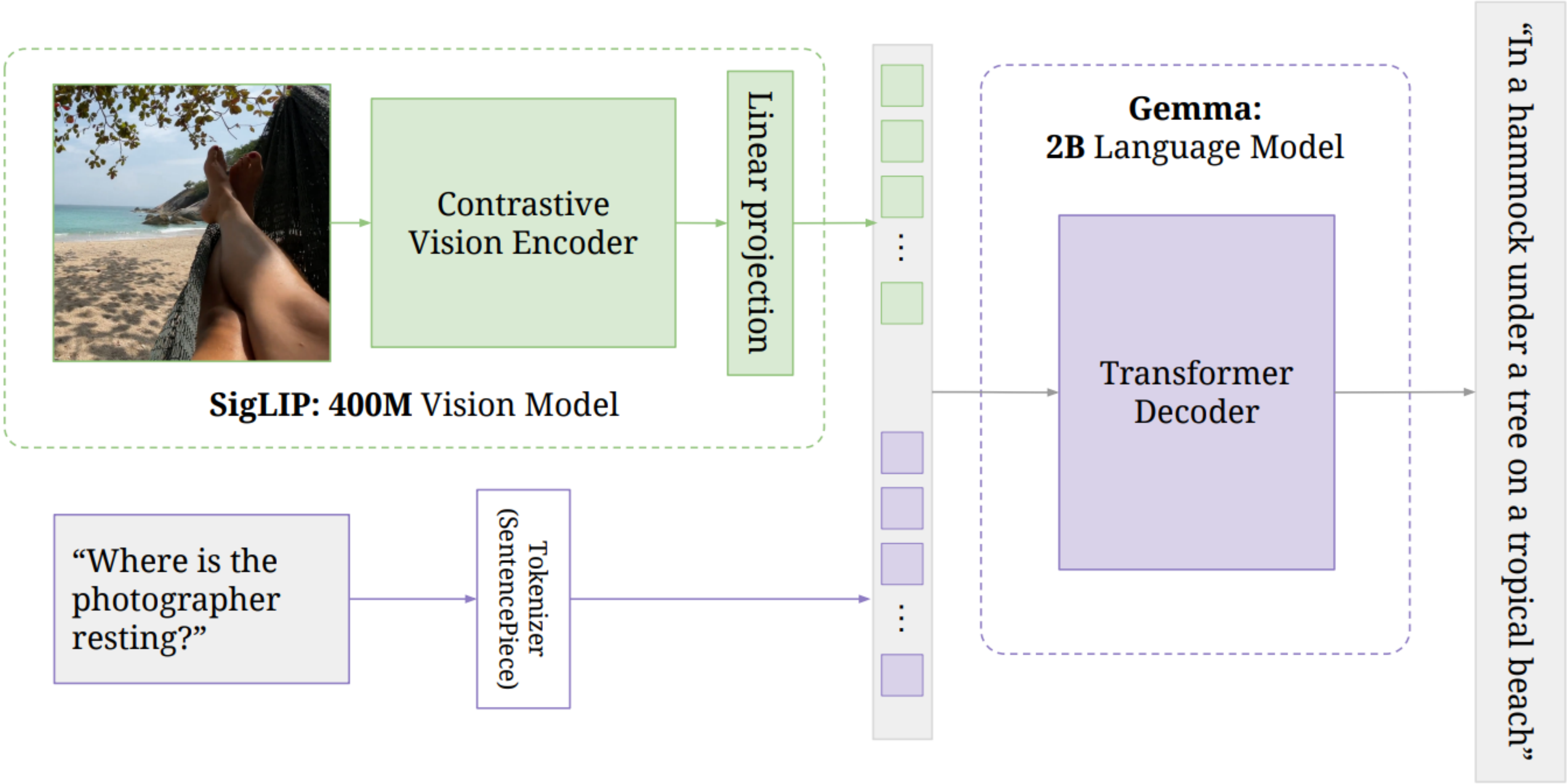


Training on GTA5 with ResNet-50 backbone and DeepLab v3+

Results



Generative VLMs

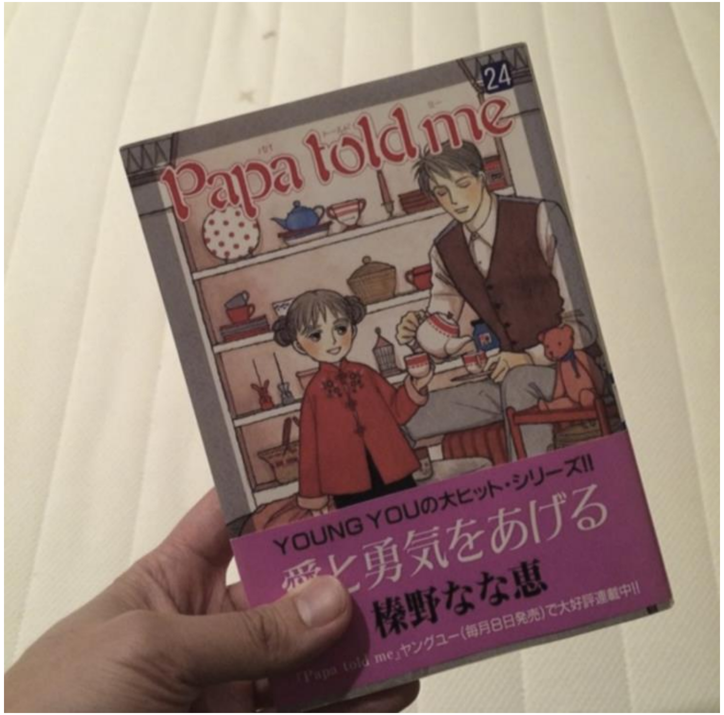


Multiple Choice Learning of Low Rank Adapters for Language Modeling

Victor Letzelter, Hugo Malard, Mathieu Fontaine, Gaël Richard, Slim Essid, Andrei Bursuc, Patrick Pérez

Technical report 2025

Handling ambiguity in (multimodal) Language Modeling



(Causal) Language Modeling

- Vocabulary $\mathcal{V} = \{1, \ldots, |\mathcal{V}|\}$
- Sequence of tokens $x \triangleq (x_t)_{t=1}^{T} \in \mathcal{V}^T$
- (In VLMs, ALMs) Context vector embeddings $c \triangleq (c_t)_{t=1}^{\tau} \in \mathbb{R}^{\tau \times d}$

Training. "Next-token-prediction" loss with "teacher-forcing"

$$\mathcal{L}(\theta) = \mathbb{E}_{c,x}[-\log p_\theta(x \mid c)] = \mathbb{E}_{c,x}\left[-\sum_{t=1}^{T} \log p_\theta(x_t \mid x_{<t}, c)\right]$$

Where the predictions can be computed *in parallel* **through causal masked attention.**

Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. Neural computation, 1989.
Vaswani et al., Attention is all you need, NeurIPS 2017

Low Rank Adaptation

(Main) Hyper-Parameters:

- rank
- which W are concerned ? e.g., Q_proj, K_proj, V_proj, Upside, Downside

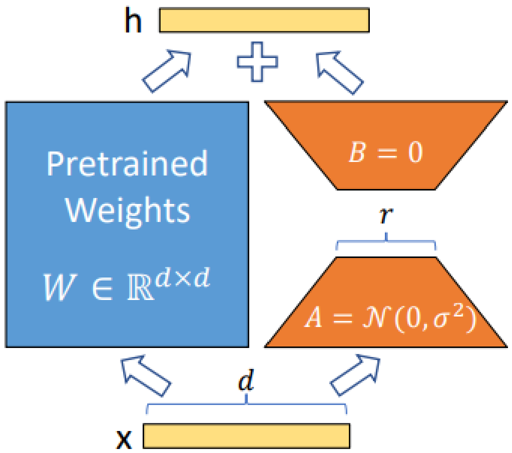


Figure 1: Our reparametrization. We only train A and B.

Hu, Edward J., Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. "Lora: Low-rank adaptation of large language models." ICLR (2022).

(Causal) Language Modeling

Inference / Decoding. Inference is performed in an autoregressive manner.

(i) Compute $p_\theta(x_1 \mid c)$ $\hat{x}_1$ select

(ii) for $t \geq 2$, compute $p_\theta(x_t \mid \hat{x}_{<t}, c)$ $\hat{x}_t$ and select

How to select the tokens ?

Maximum A Posteriori Techniques (greedy, beam search, diverse beam search) aim at maximizing the prob $p_\theta(\hat{x})$ of the predicted sequence

(Truncated) Sampling (top-k, nucleus) allow to increase the diversity.

-> Quality / Diversity trade-off.

While some decoding methods allow to improve the test-time diversity, we aim at **learning** to produce diverse outputs.

Multiple Choice Learning to Language Modeling

Winner-takes-all optimization

1. For each training sample (c, x) in the batch $\mathcal{B}$: Compute $p(x \mid c; \theta_k)$ for $k \in \{1, \ldots, K\}$, and choose the best model $k^\star(x, c) = \operatorname{argmax}_k p(x \mid c; \theta_k)$.
2. Compute the winner-takes-all (WTA) loss as:

$$\mathcal{L}^{\mathrm{WTA}}(\theta_1, \ldots, \theta_K) = -\mathbb{E}_{c,x}\left[\max_{k=1,\ldots,K} \log p(x \mid c; \theta_k)\right], \tag{2}$$

where $\log p(x \mid c; \theta_k) = \sum_{t=1}^{T} \log p(x_t \mid x_{<t}, c; \theta_k)$, and perform an optimization step.

Multiple Choice Learning to Language Modeling

Relaxed WTA objective

$$\mathcal{L}^{\mathrm{WTA}}(\theta) = -\mathbb{E}_{c,x}\left[\sum_{k=1}^{K} q_k \log p(x \mid c; \theta_k)\right]$$

$$q_k = \tfrac{\varepsilon}{K-1} \text{ for } k \neq k^\star$$

C. Rupprecht et al., Learning in an uncertain world: Representing ambiguity through multiple hypotheses, ICCV 2017

MCL with Multiple Low Rank Adapters

Let θ be the parameters of the pretrained base model. At each layer ℓ where LoRA is enabled, we use a family of adapters $(A_\ell^k, B_\ell^k) \in \mathbb{R}^{d\times r} \times \mathbb{R}^{r\times d}$ for $k \in \{1, \ldots, K\}$. Let

$$\theta_k = \theta \cup \left\{(A_\ell^k, B_\ell^k) \mid \ell = 1, \ldots, L\right\} \ ,$$

LoRA units at layer ℓ are then computed as:

$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_K \end{bmatrix} \leftarrow \begin{bmatrix} B_\ell^1 A_\ell^1 & 0 & 0 & 0 \\ 0 & B_\ell^2 A_\ell^2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & B_\ell^K A_\ell^K \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_K \end{bmatrix} + \begin{bmatrix} f_\theta^\ell(\mathbf{x}_1) \\ f_\theta^\ell(\mathbf{x}_2) \\ \vdots \\ f_\theta^\ell(\mathbf{x}_K) \end{bmatrix},$$

Experiments

Audio Captioning

- Qwen-2-Audio. ~8B params $|\mathcal{V}| = 156{,}032.$
- Datasets: Clotho-V2, AudioCaps

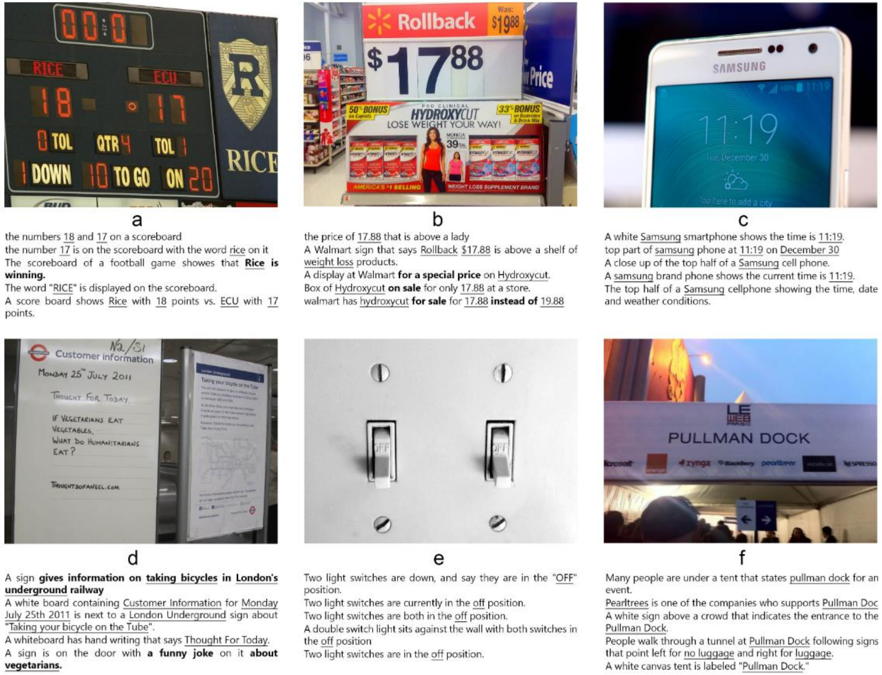
Image Captioning

- LLaVA 1.6. ~7B params $|\mathcal{V}| = 32{,}000.$
- Datasets: TextCaps

Metrics

- Test-loss (or perplexity),
- Natural language generation quality metrics (BLEU-n, ROUGE, METEOR)
- Captioning metrics CIDEr, SPICE and SPIDEr

- Perceptual metrics: Sentence-BERT (sBERT)
- Diversity metrics (Div-1, Div-2, mBLEU-4)



Y. Chu, Yunfei et al., Qwen2-audio technical report, arXiv 2024
H. Liu et al., Visual instruction tuning, NeurIPS 2023

Experiments

Baselines

- Next-token-prediction baseline (1 hyp).
- Decoding: Greedy, Beam Search, Diverse Beam Search.

  -> Alignment of the number of trainable parameters and number of forward passes at inference.

  -> The runs are otherwise perfectly comparable (same architecture, learning rate, training details, random seed).

Results - audio captioning

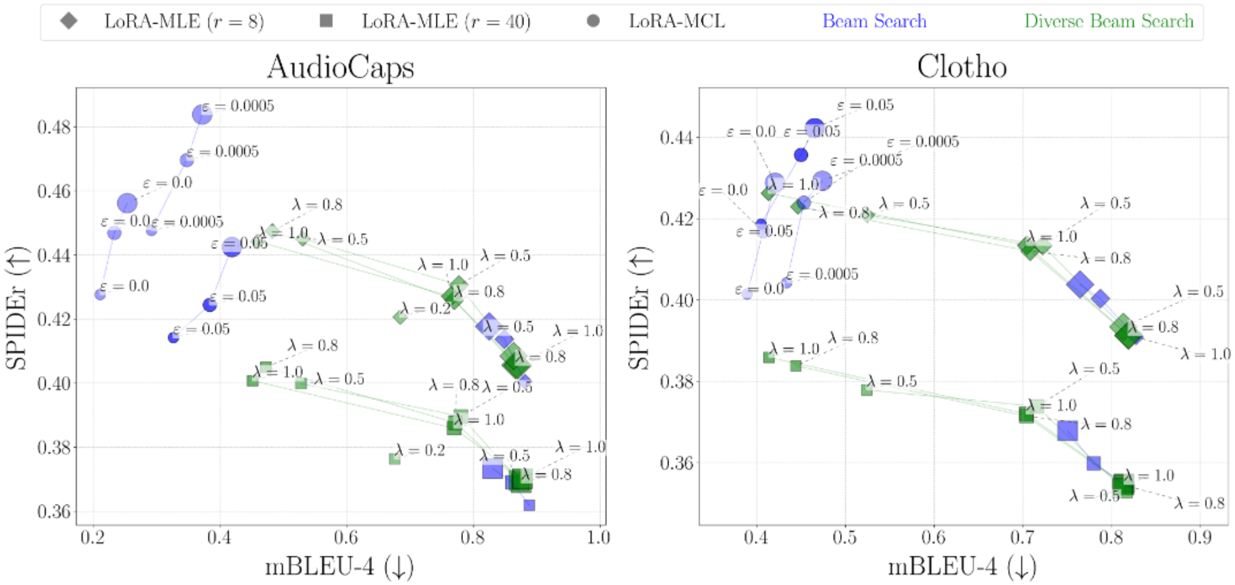


Figure 2: **Quality vs. Diversity on Audio Captioning with 5 candidates.** Quality measured by SPIDEr (↑) and Diversity by mBLEU-4 (↓). Marker shape stands for the method, and size is proportional to the number of forwards performed per example at inference time. Note that `LoRA-MLE` was trained with two rank values, $r = 8$ and $r = 8K$, for fair comparison in terms of parameter count. Color corresponds to the decoding method (beam search or diverse beam search). Values of ε and λ for `LoRA-MCL` and diverse beam search, resp., are indicated in the plot, with the color shade proportional to the corresponding parameter value to better distinguish the markers.

Results - image captioning

Table 2: **Quality and Diversity Evaluation on TextCaps with 3 candidates.** Best scores are in **bold**, second-best are underlined. For each of the presented metrics, higher is better (↑) except for mBLEU-4 (↓). `LoRA-MCL` is trained with $\varepsilon = 0.1$, $r = 8$ and $\alpha = 32$. LoRA-MLE is trained with $r = 24$ and $\alpha = 4 \times r = 96$, ensuring the same dynamics and number of parameters across models.

Training	Decoding	Beam	mBLEU-4	BLEU-4	METEOR	sBERT	CIDEr-D	SPICE	SPIDEr
LoRA-MLE	Beam Search	3	0.688	0.318	0.315	0.670	1.517	0.244	0.873
LoRA-MLE	Beam Search	6	0.786	0.338	0.326	0.671	1.557	0.246	0.895
LoRA-MLE	DBS ($\lambda = 0.8$)	3	<u>0.437</u>	<u>0.349</u>	0.327	0.686	1.590	0.251	0.909
LoRA-MLE	DBS ($\lambda = 1.0$)	3	**0.416**	0.348	0.326	0.685	1.586	0.250	0.906
LoRA-MLE	DBS ($\lambda = 0.8$)	6	0.671	0.341	0.328	0.681	1.573	0.251	0.903
LoRA-MLE	DBS ($\lambda = 1.0$)	6	0.666	0.340	0.328	0.680	1.577	0.250	0.904
LoRA-MCL	Greedy	1	0.520	0.344	<u>0.330</u>	**0.690**	**1.674**	<u>0.255</u>	**0.955**
LoRA-MCL	Beam Search	2	0.490	**0.360**	**0.333**	<u>0.687</u>	<u>1.627</u>	**0.258**	<u>0.932</u>

Hypotheses specialization

Table 3: **SPIDEr (↑) & mBLEU-4 (↓) on different parts of synthetic test set.**

Test subset	Training	SPIDEr	mBLEU-4
French	LoRA-MLE	0.411	0.138
	LoRA-MCL	**0.464**	**0.027**
English	LoRA-MLE	**0.756**	0.126
	LoRA-MCL	0.722	**0.029**

LoRA-MLE.
{A bottle of Cerveza is on a table.}
{Une bouteille de vin de cidre de cidre de cidre [...]}
LoRA-MCL.
{A bottle of beer with a label that says "Sel Maguet"}
{Une bouteille de vin est étiquetée avec le mot « Maguay ».}

LoRA-MLE.
{A book titled Papa Told Me is being held by a person.}
{A book called Papa told me is being held by a person.}
Lora-MCL.
{A book titled Papa Told Me is being held by a person}
{Un livre papier intitulé Papa Told Me.}

Figure 3: **Observing specialization in bilingual image description.** Quantitative (*Left*) and Qualitative (*Right*) analysis for LoRA-MLE and LoRA-MCL in the setup of Section 5.3.2.

Specialization observed: The winning head is the first one in ~ 89% of the French captions and the second one in ~ 97% of the English captions

Conclusion and perspectives

Conclusion

- Stage 2 in foundation models has diversified significantly
  - Multiple opportunities for (relatively) low-cost adaptations, improvements, studies

- Language steps into multiple traditional computer vision tasks and communities
  - But this is not the end of the story

- FM know a lot, but they don't know it all: uncertainty quantification more relevant than ever

---

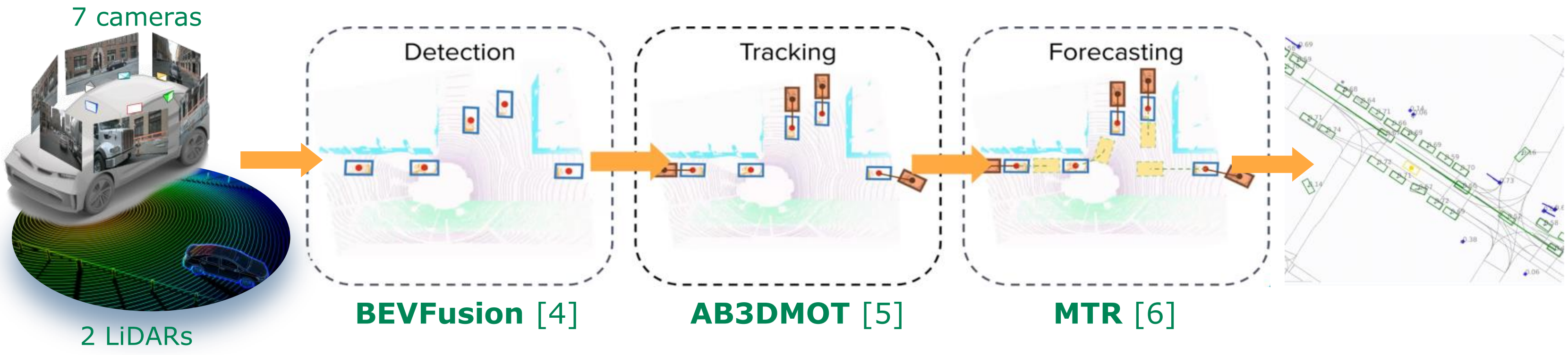
Perspectives

- Language and VLMs/MLLMs bring new sources of uncertainty to consider:
  - **Prompt corruption:** reordering, rephrasings, perturbations
  - **Task ambiguity:** *What is happening in this image?*
  - **Knowledge gaps & training coverage:** dataset cut-off date
  - **Prompt underspecification:** Who is the president?
  - **Reasoning complexity & compositionality:** errors in intermediate steps, complex vision-text tasks
  - **Multimodal grounding errors:** hallucination on the LLM side
  - **Decoding randomness:** stochastic decoding

G. Franchi et al., Technical report coming soon!

Perspectives

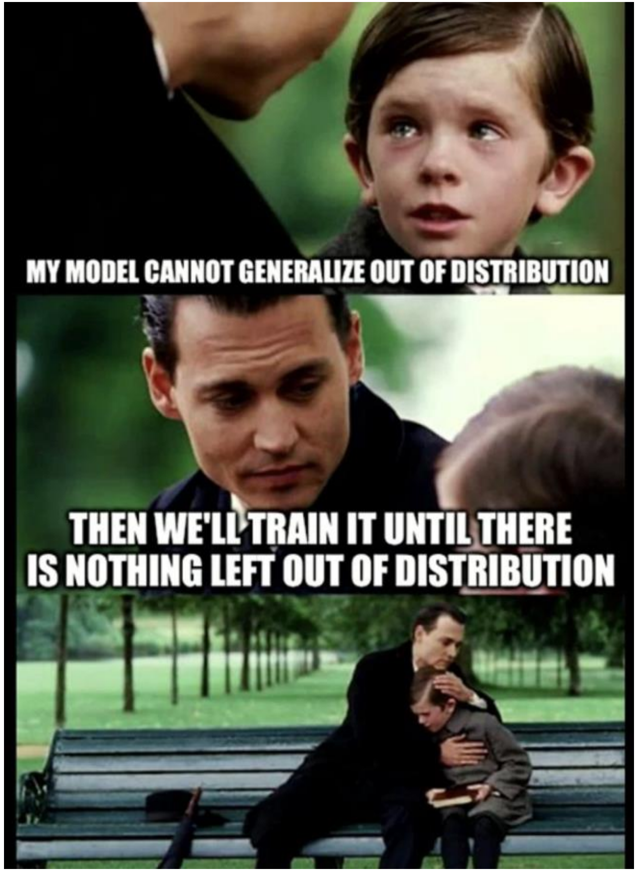
- Reliability of embodied systems: end-to-end autonomous driving, robots, etc.
    - Errors accumulating from intermediate modules
    - Errors accumulating in time.



Y. Xu et al., How to perform end-to-end forecasting? Valeo4Cast — A Modular Approach, ECCV ROAD++ Workshop 2024

Perspectives

- While revisiting many practices, proper evaluation is as critical as ever
- How to evaluate models trained on different datasets?
- The pretraining distribution of most foundation models is undisclosed
  - It might be your test distribution (:



Conclusion and perspectives

- Stage 2 in foundation models has diversified significantly
- Language steps in
- FM know a lot, but they don’t know it all

- Language and VLMs/MLLMs bring new sources of uncertainty to consider
- A lot to do for reliability of embodied systems: end-to-end autonomous driving, robots, etc.
- New evaluation strategies needed

https://valeoai.github.io