[1] Inria
[2] valeo.ai
[3] kyutai

Code & Demo
https://astra-vision.github.io/FAMix/

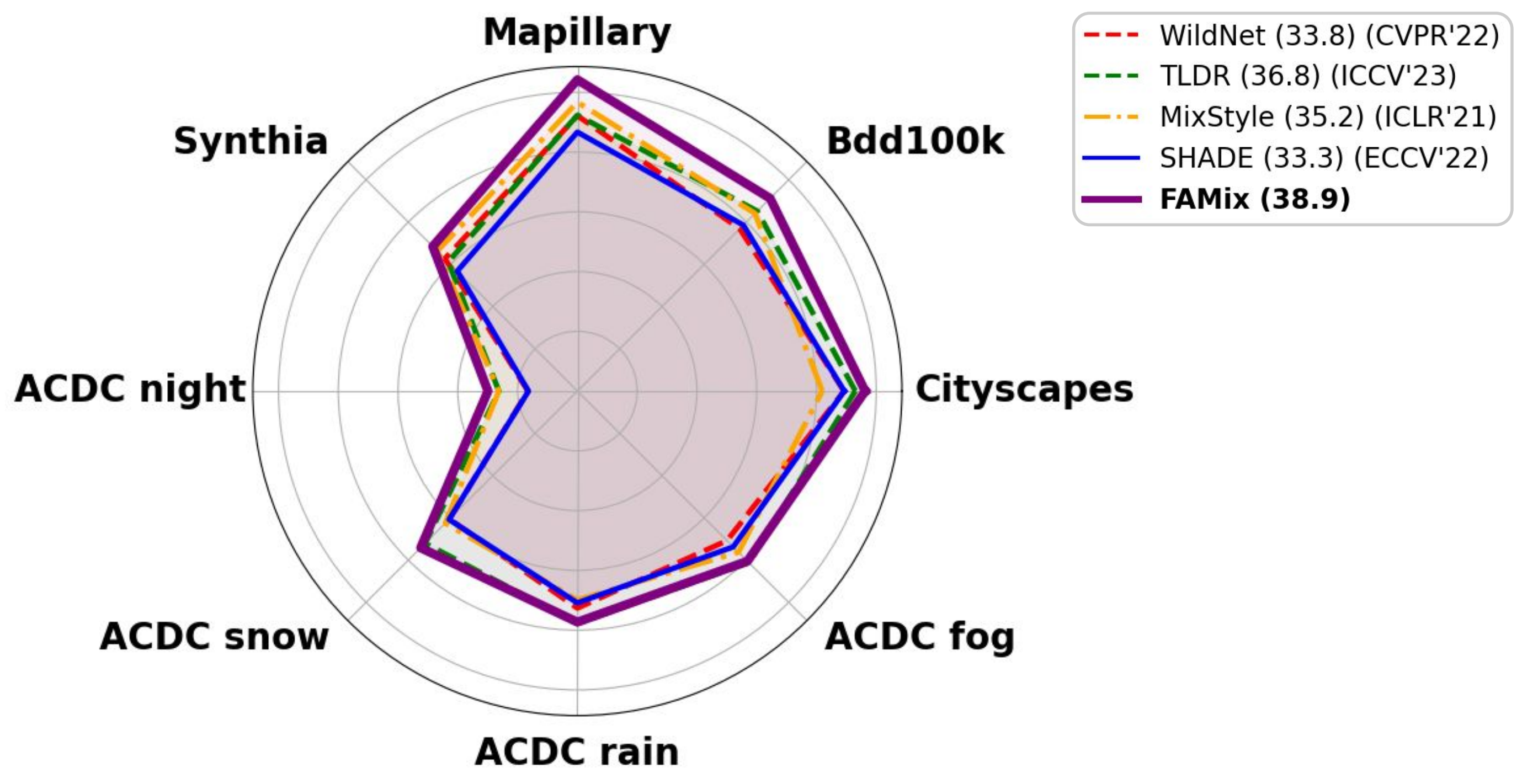
A Simple Recipe for Language-guided Domain Generalized Segmentation

Mohammad Fahes[1], Tuan-Hung Vu[1,2], Andrei Bursuc[1,2], Patrick Pérez[3], Raoul de Charette[1]

CVPR SEATTLE, WA JUNE 17-21, 2024

What is FAMix?

A recipe for Domain Generalized Semantic Segmentation using CLIP pretraining.

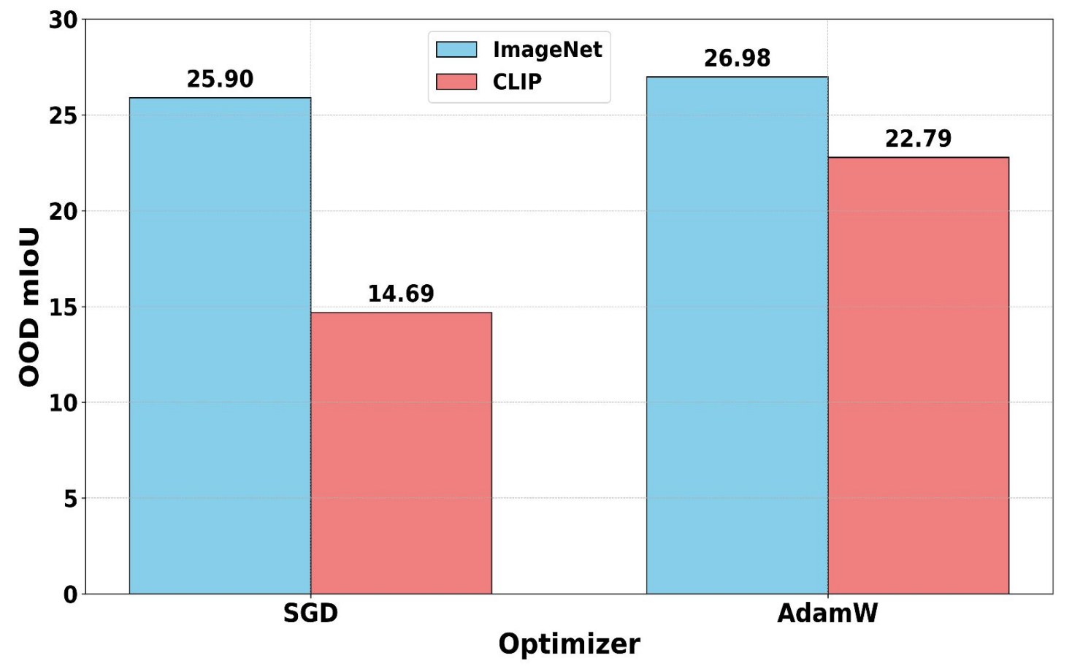


Training on GTA5 with ResNet-50 backbone and DeepLab v3+

WildNet & TLDR use extra images.
MixStyle is applied with CLIP and our minimal fine-tuning component.

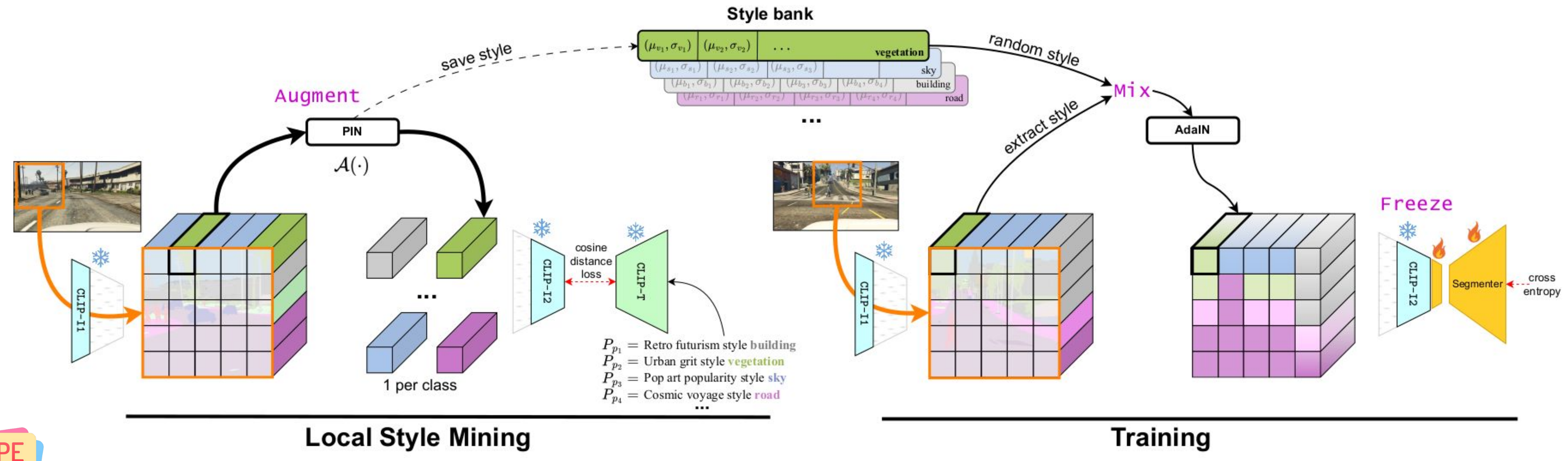
Motivation

CLIP exhibits **distributional robustness** and offers the **language modality** which can help visual task.



Naive end-to-end fine-tuning is not satisfying!

How to use CLIP for enhanced Domain Generalization in Semantic Segmentation?



P_{p_1} = Retro futurism style building
P_{p_2} = Urban grit style vegetation
P_{p_3} = Pop art popularity style sky
P_{p_4} = Cosmic voyage style road
...

RECIPE

Augment

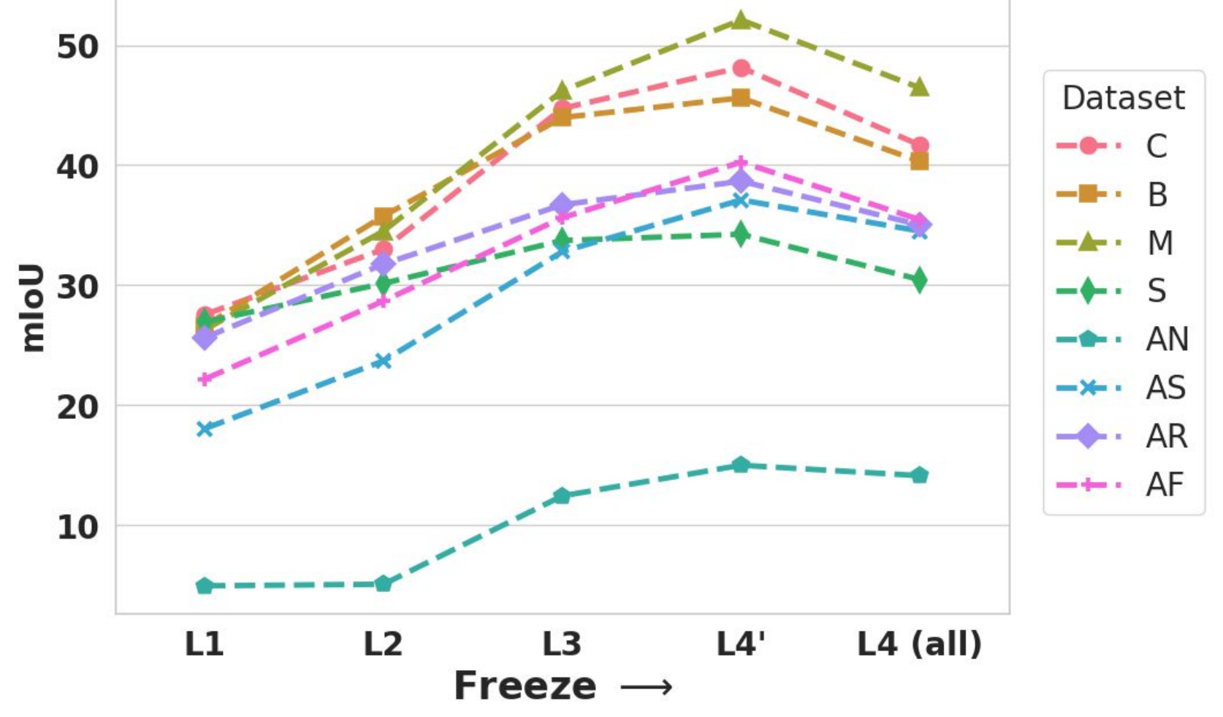
random prompt — feature patch

$$\mathrm{PIN}_{(P_p)}(f_p) = \sigma \frac{f_p - \mu(f_p)}{\sigma(f_p)} + \mu$$

Optimize with *Gradient Descent* by bringing embeddings of f_p and P_p closer

Freeze

Fine-tune only last few layers!



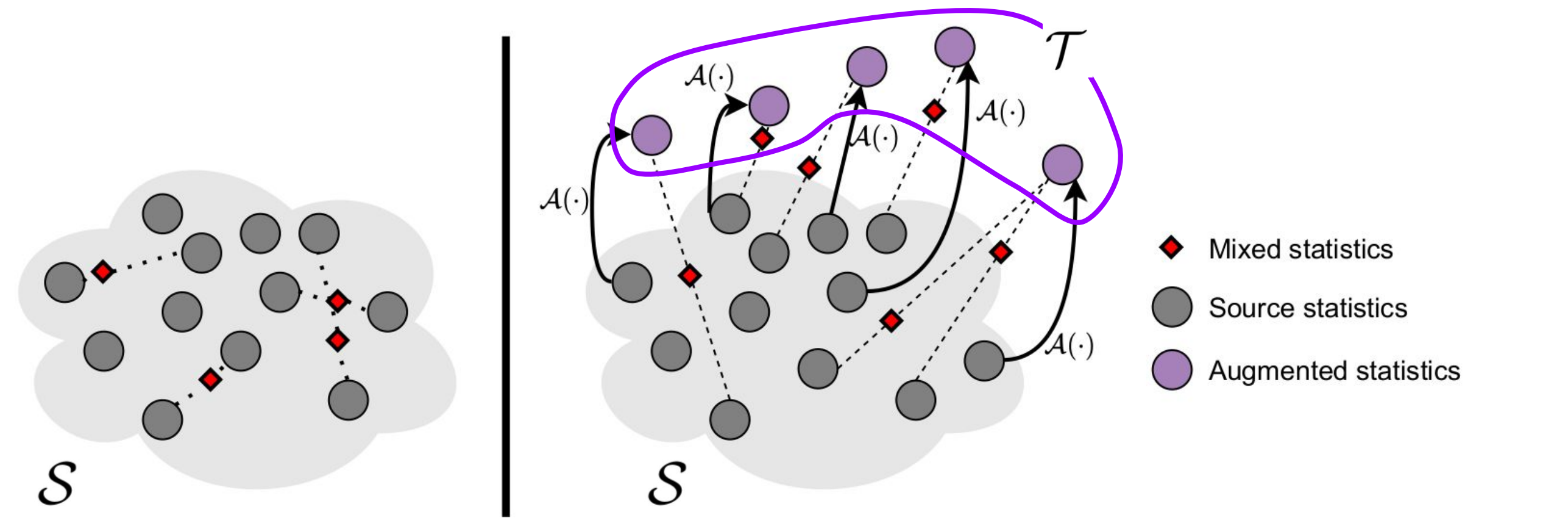
Mix

Sampled from $\mathcal{T}$

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

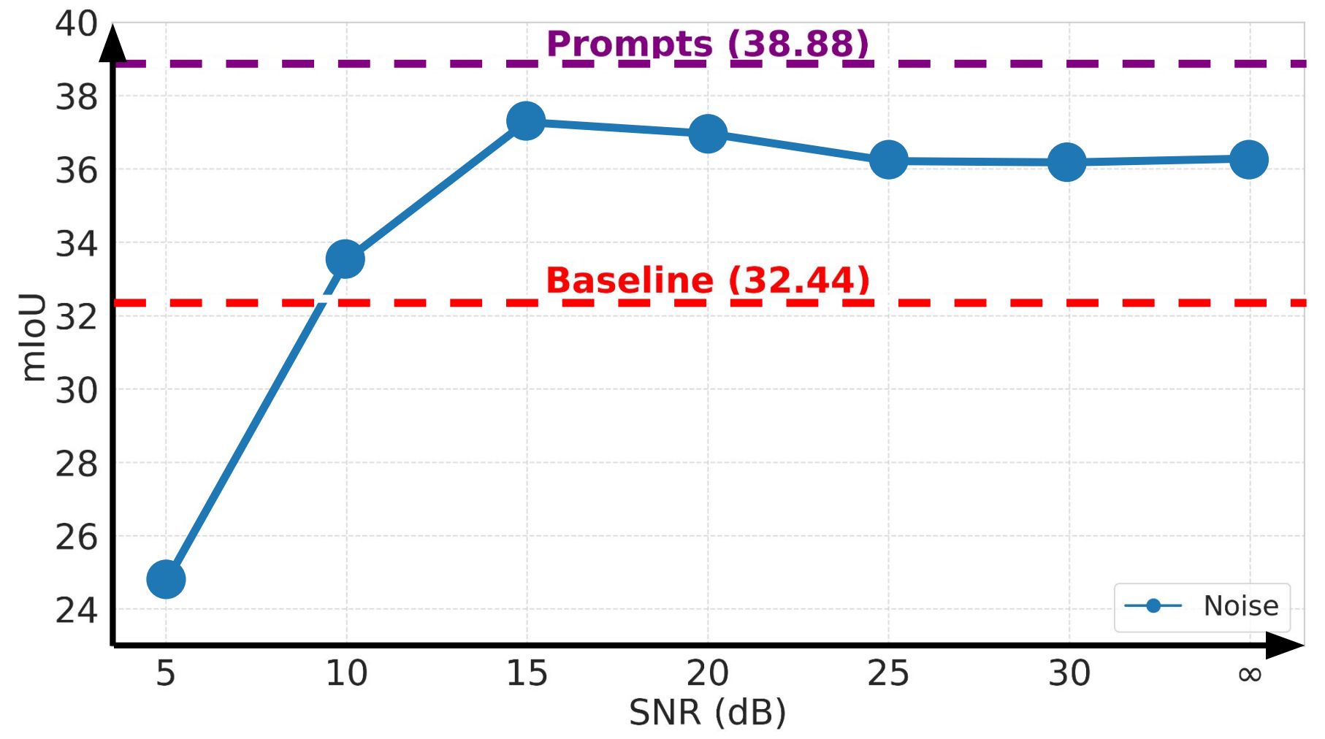
$\alpha \in \mathbb{R}^C$

Multiplications are element-wise



FAMix with $\mathcal{S}$ = 36.29 | $\mathcal{S} \cup \mathcal{T}$ = 37.64 | $\mathcal{T}$ = **38.88**

Language or Noise?



Prompt-driven augmentation > noise augmentation

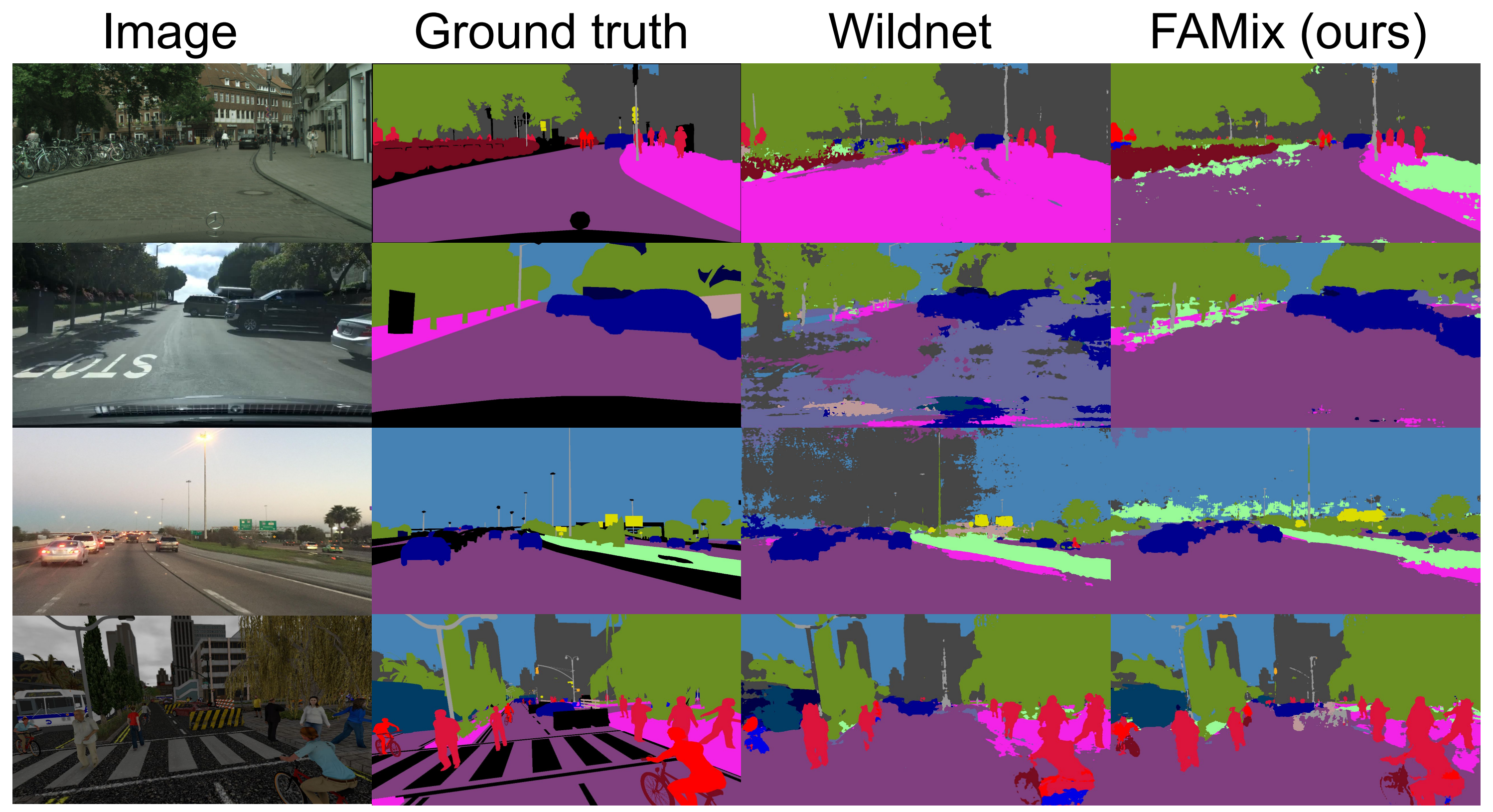
Prompt construction?

e.g. *"wqvsecpas style"* → Random Characters; e.g. *"Galactic Fantasy style"* → Random Style; e.g. *"road"* → Class Name

Random Characters	Random Style	Class Name	C	B	M	S	AN	AS	AR	AF	Mean
		✓	45.99	43.71	50.48	**34.75**	15.22	35.09	34.92	38.17	37.29
✓			46.10	44.24	48.90	33.62	13.39	35.99	36.68	39.86	37.35
	✓		45.64	44.59	49.13	33.64	**15.33**	**37.32**	35.98	38.85	37.56
✓		✓	47.83	44.83	50.38	34.27	14.43	37.07	37.07	38.76	38.08
	✓	✓	**48.15**	**45.61**	**52.11**	34.23	14.96	37.09	**38.66**	**40.25**	**38.88**

Any prompt improves generalization in our framework!

Qualitative results



Removing one ingredient spoils the recipe!

Freeze	Augment	Mix	C	B	M	S	AN	AS	AR	AF	Mean
✗	✗	✗	16.81	16.31	17.80	27.10	2.95	8.58	14.35	13.61	14.69
✗	✓	✗	22.48	26.05	24.15	25.40	4.83	17.61	22.86	19.75	20.39
✗	✗	✓	20.07	21.24	22.91	26.52	1.28	14.99	22.09	20.51	18.70
✗	✓	✓	27.53	26.59	26.27	26.91	4.90	18.91	25.60	22.14	22.36
✓	✗	✗	37.83	38.88	44.24	31.93	12.41	29.59	31.56	33.05	32.44
✓	✓	✗	36.65	35.73	37.32	30.44	14.72	34.65	34.91	38.98	32.93
✓	✗	✓	43.43	43.79	48.19	33.70	11.32	35.55	36.15	38.19	36.29
✓	✓	✓	**48.15**	**45.61**	**52.11**	**34.23**	**14.96**	**37.09**	**38.66**	**40.25**	**38.88**

Takeaways

- ➔ One can **"mine" random styles** using random language prompts and feature patches
- ➔ **Training only the last layers** preserves CLIP robustness and helps adapting to the segmentation task
- ➔ Applying **style mixing between original and augmented statistics** significantly outperforms MixStyle on single source domain generalization

WildNet, Lee et al., CVPR 22
SHADE, Zhao et al., ECCV 22
TLDR, Kim et al., ICCV 23
MixStyle, Zhou et al., ICLR 21
PØDA, Fahes et al., ICCV 23