Decoupled Finetuning for Domain Generalizable Semantic Segmentation

Jaehyun Pahk[1], Donghyeon Kwon[1], Seong Joon Oh[3], and Suha Kwak[1 2]

Dept. of CSE, POSTECH[1], Graduate School of AI, POSTECH[2], Tübingen AI Center, Universität Tübingen[3]

Problems in Joint Finetuning

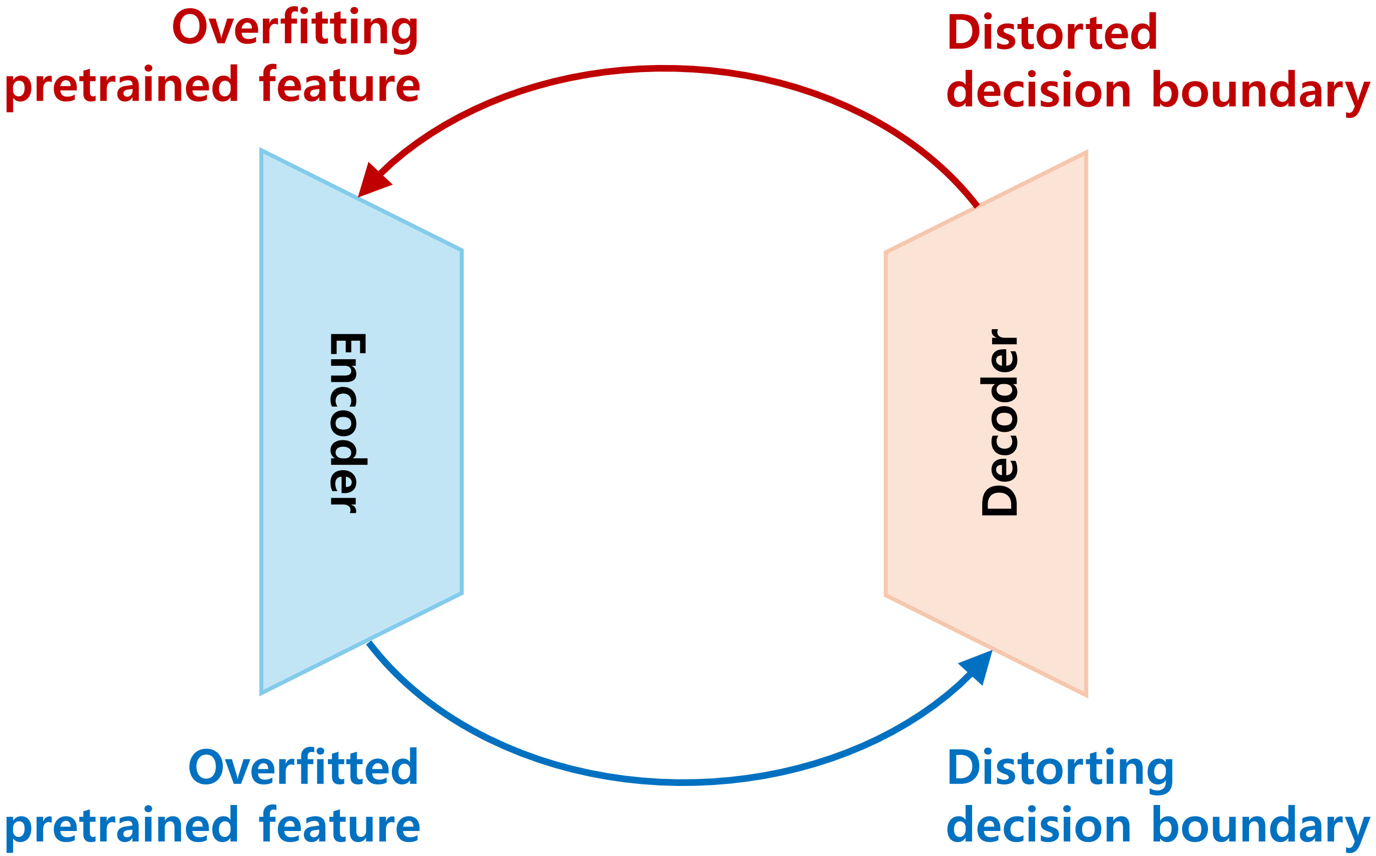
Joint finetuning of a pretrained encoder and a randomly initialized decoder ➜ *“de facto standard”*

However…

Problems in Joint Finetuning

Joint finetuning of a pretrained encoder and a randomly initialized decoder **➜** *“de facto standard”*

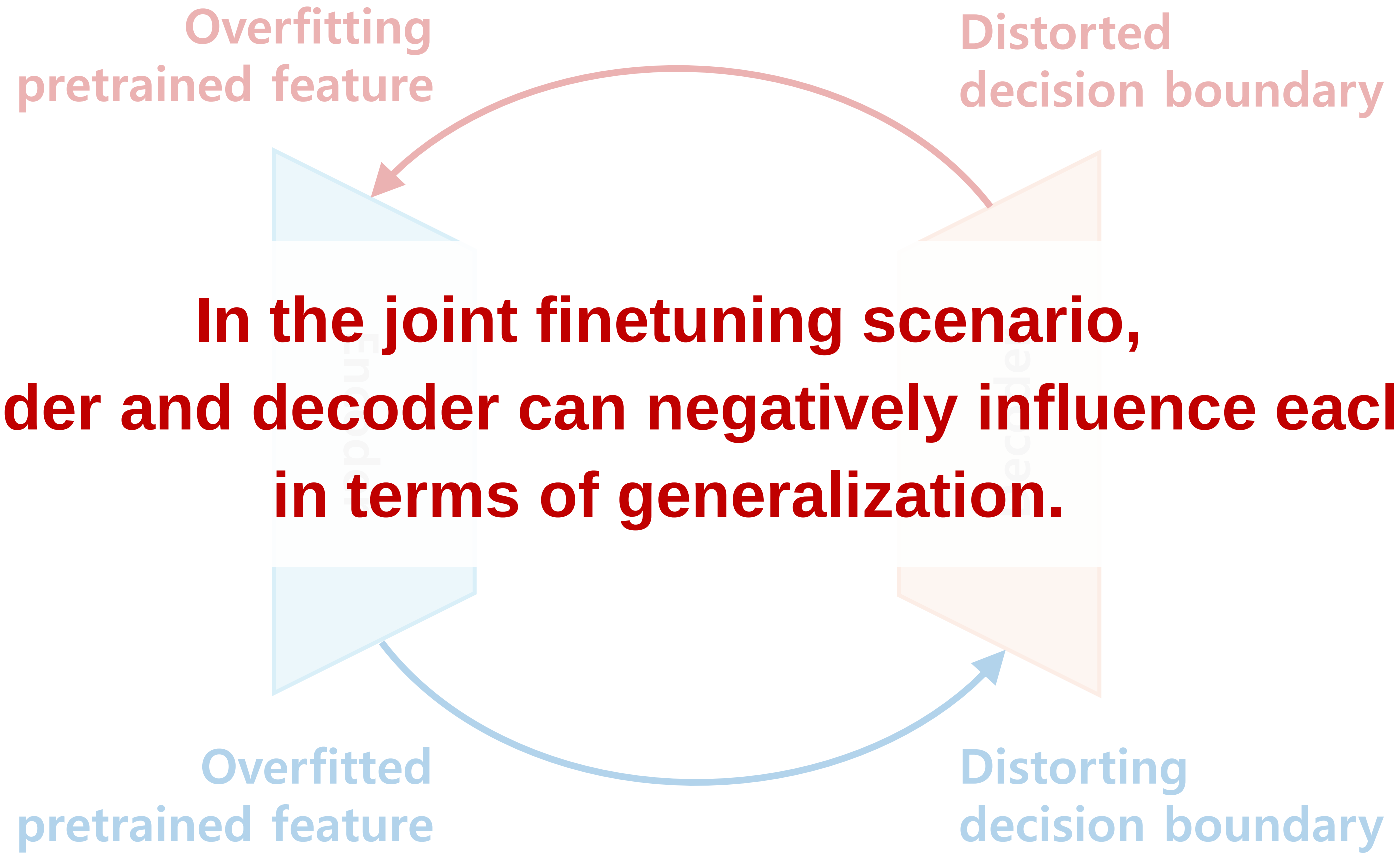
However…



Problems in Joint Finetuning

Joint finetuning of a pretrained encoder and a randomly initialized decoder ➜ *"<u>de facto standard</u>"*

However…

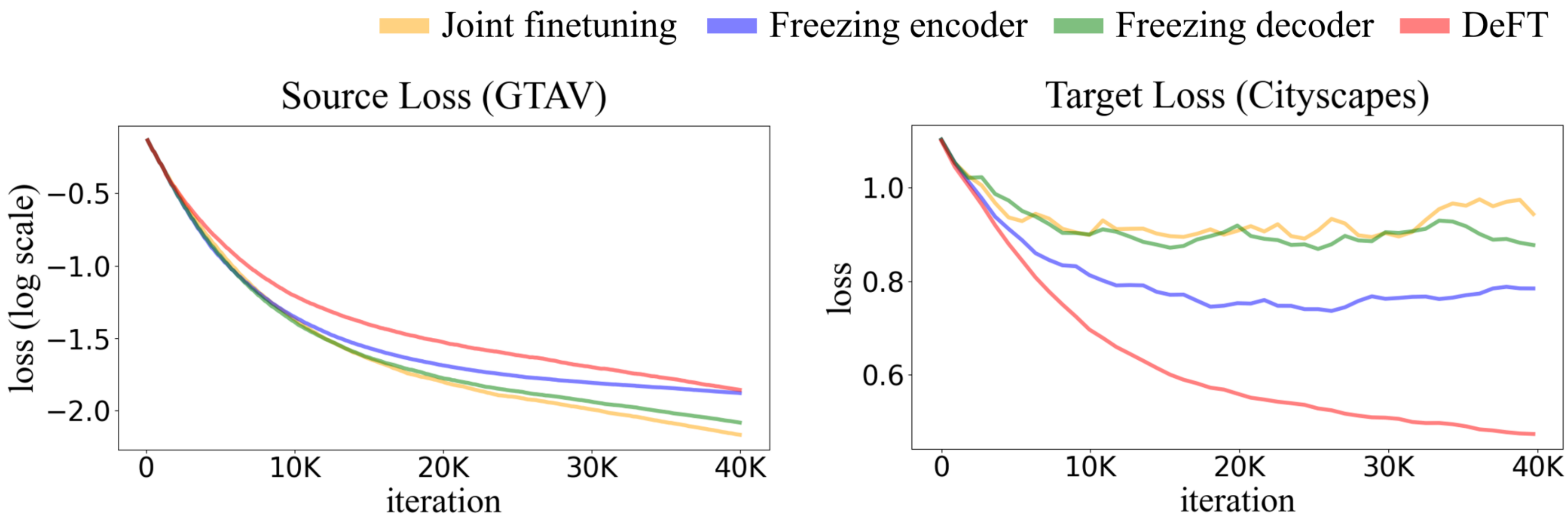


In the joint finetuning scenario,
the encoder and decoder can negatively influence each other
in terms of generalization.

Naïve Solution: Simply Freezing One Module

Simply freezing either the encoder or decoder

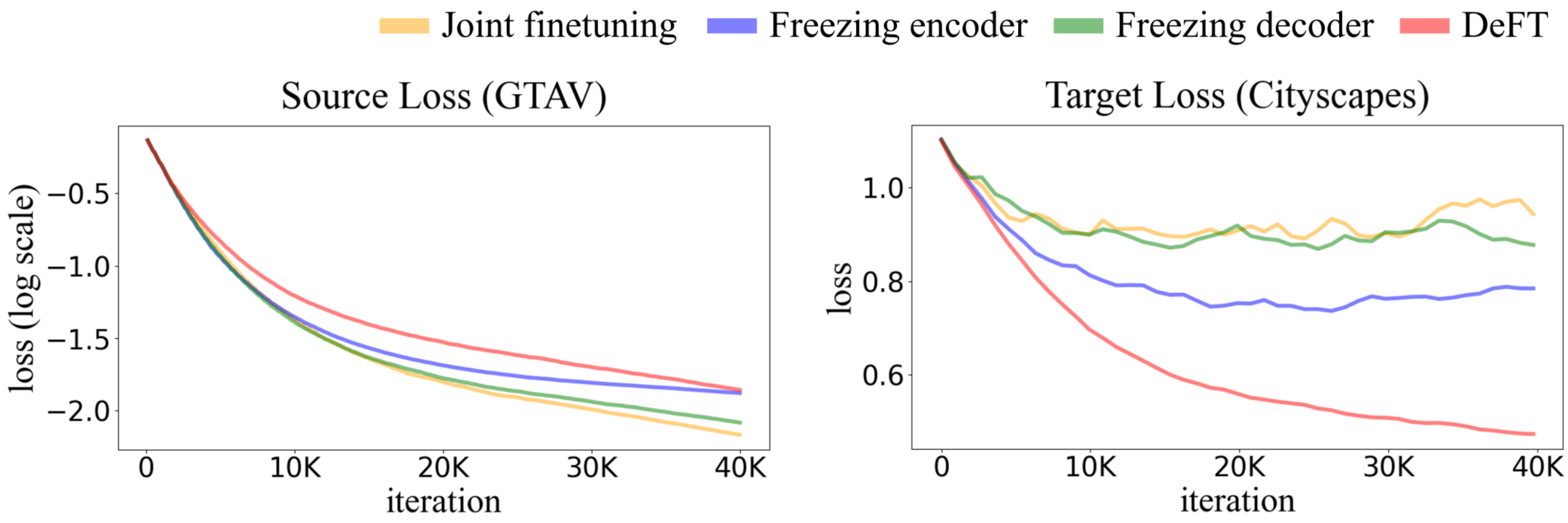
= preventing one module from being distorted by its overfitted counterpart



Naïve Solution: Simply Freezing One Module

Simply freezing either the encoder or decoder

= preventing one module from being distorted by its overfitted counterpart

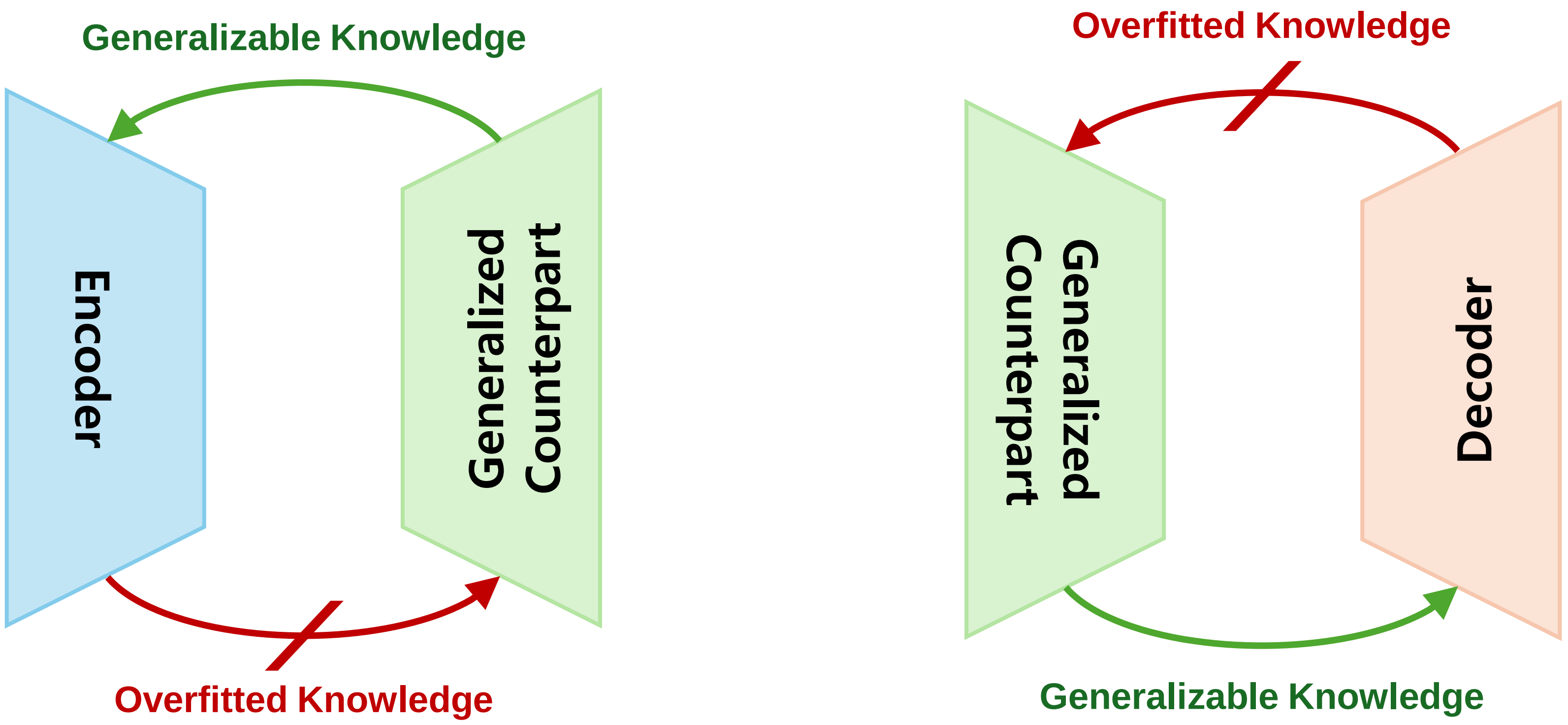


Of course, this is far from the optimal solution (*lack of task-relevant knowledge of the frozen module*)

Decoupled Finetuning (DeFT) Framework

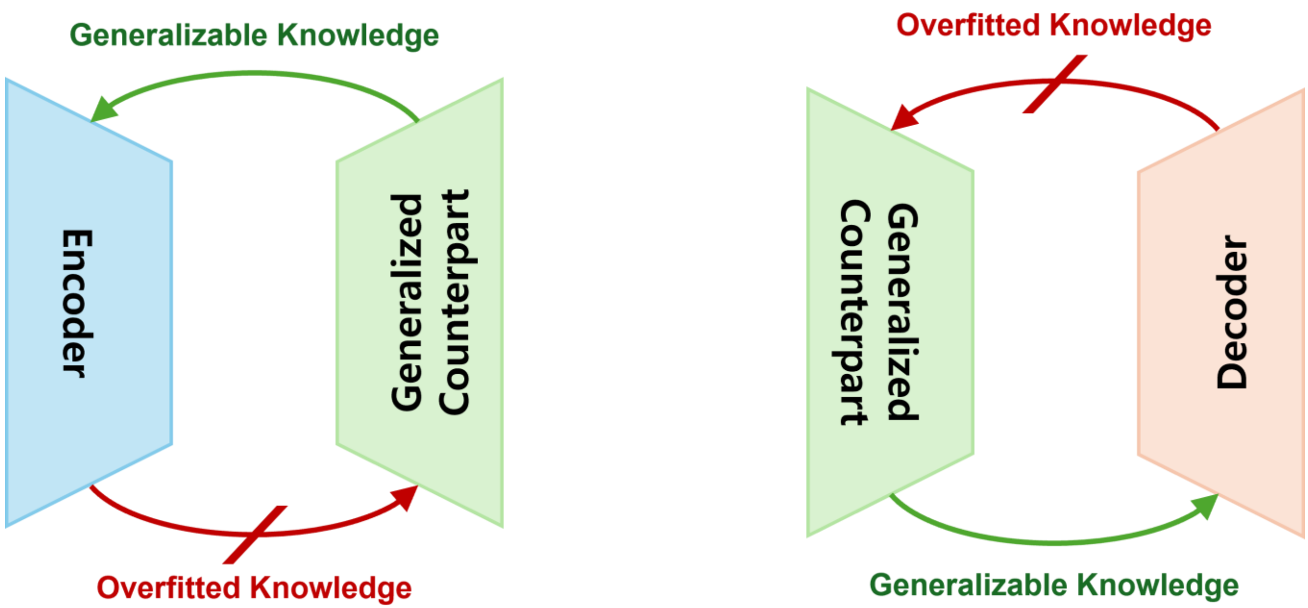
We propose ***Decoupled FineTuning (DeFT)*** – a novel training framework for generalizable segmentation

Decoupling the encoder and decoder
→ Coupling with counterparts retaining domain-generalizable knowledge during finetuning.



Decoupled Finetuning (DeFT) Framework

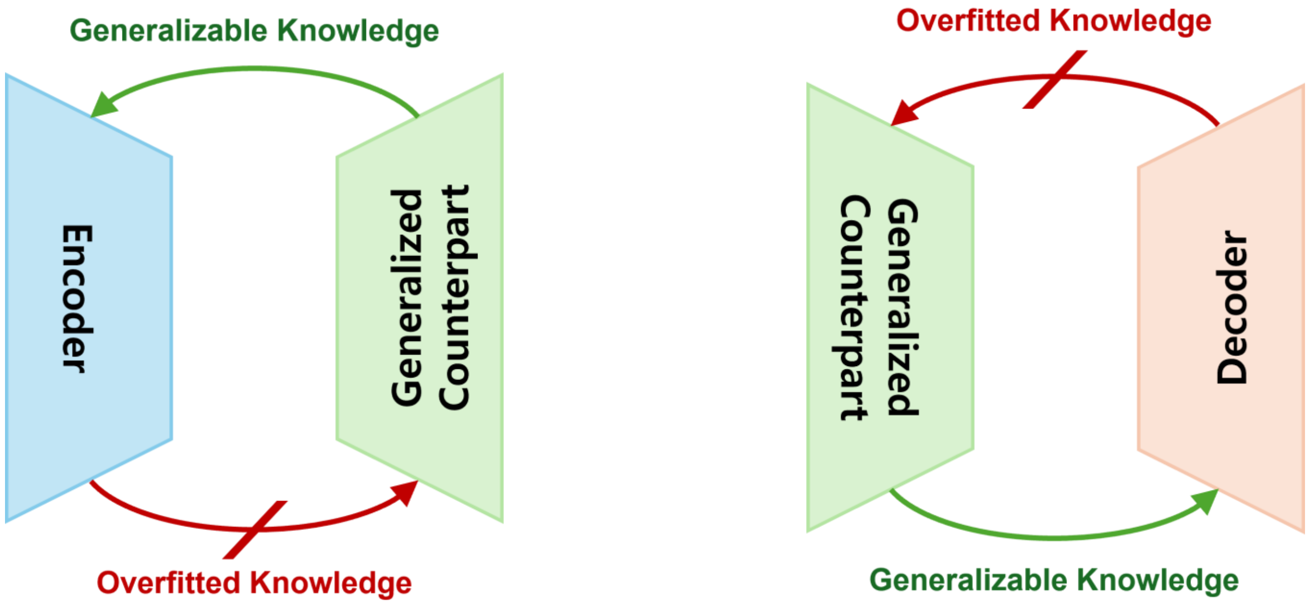
We propose ***<u>Decoupled FineTuning (DeFT)</u>*** – a novel training framework for generalizable segmentation



1. What should be used as the generalized counterpart?
2. How can the alignment of the decoupled encoder and decoder be ensured?

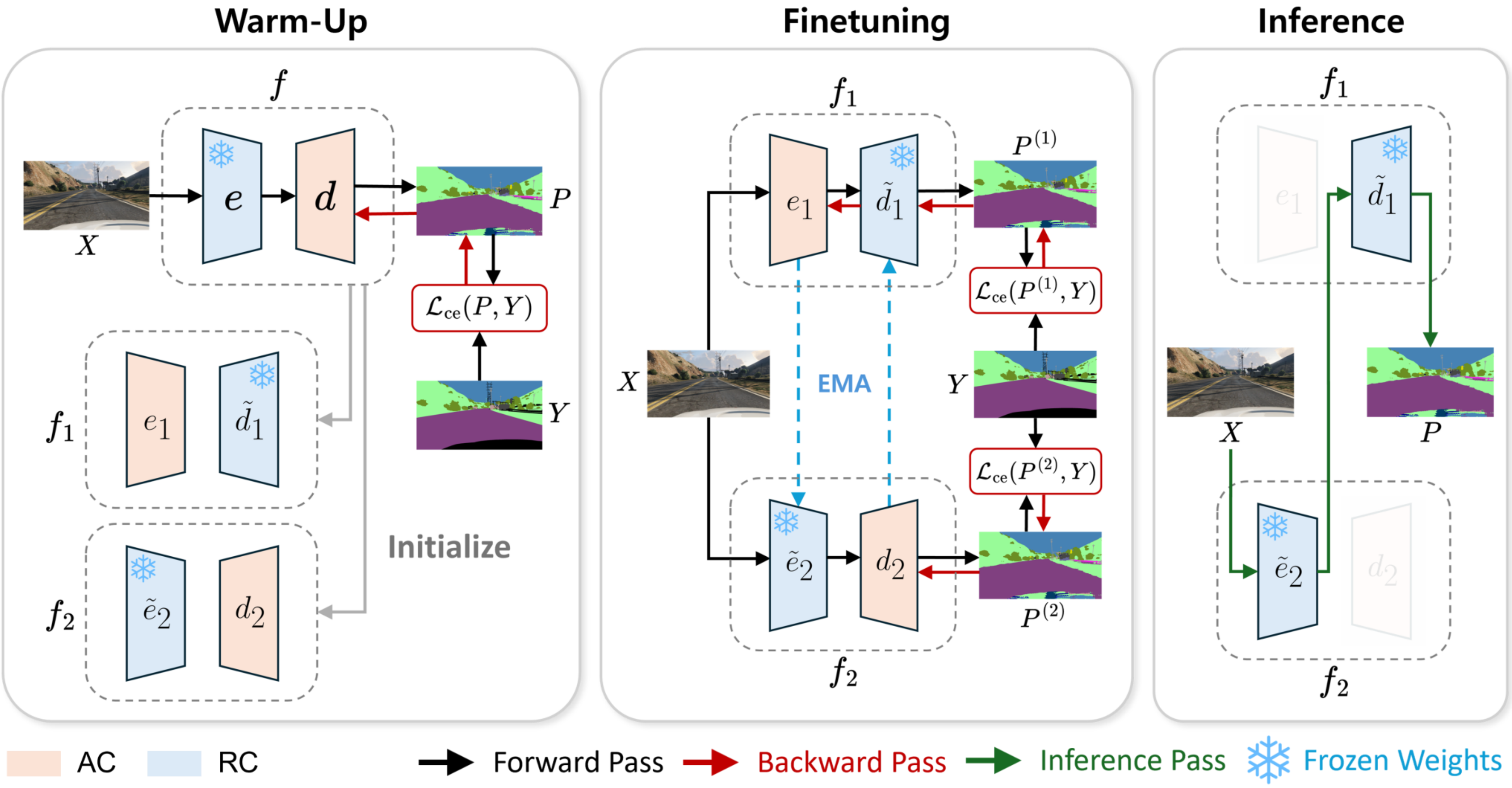
Decoupled Finetuning (DeFT) Framework

We propose ***De**coupled **F**ine**T**uning (**DeFT**)* – a novel training framework for generalizable segmentation

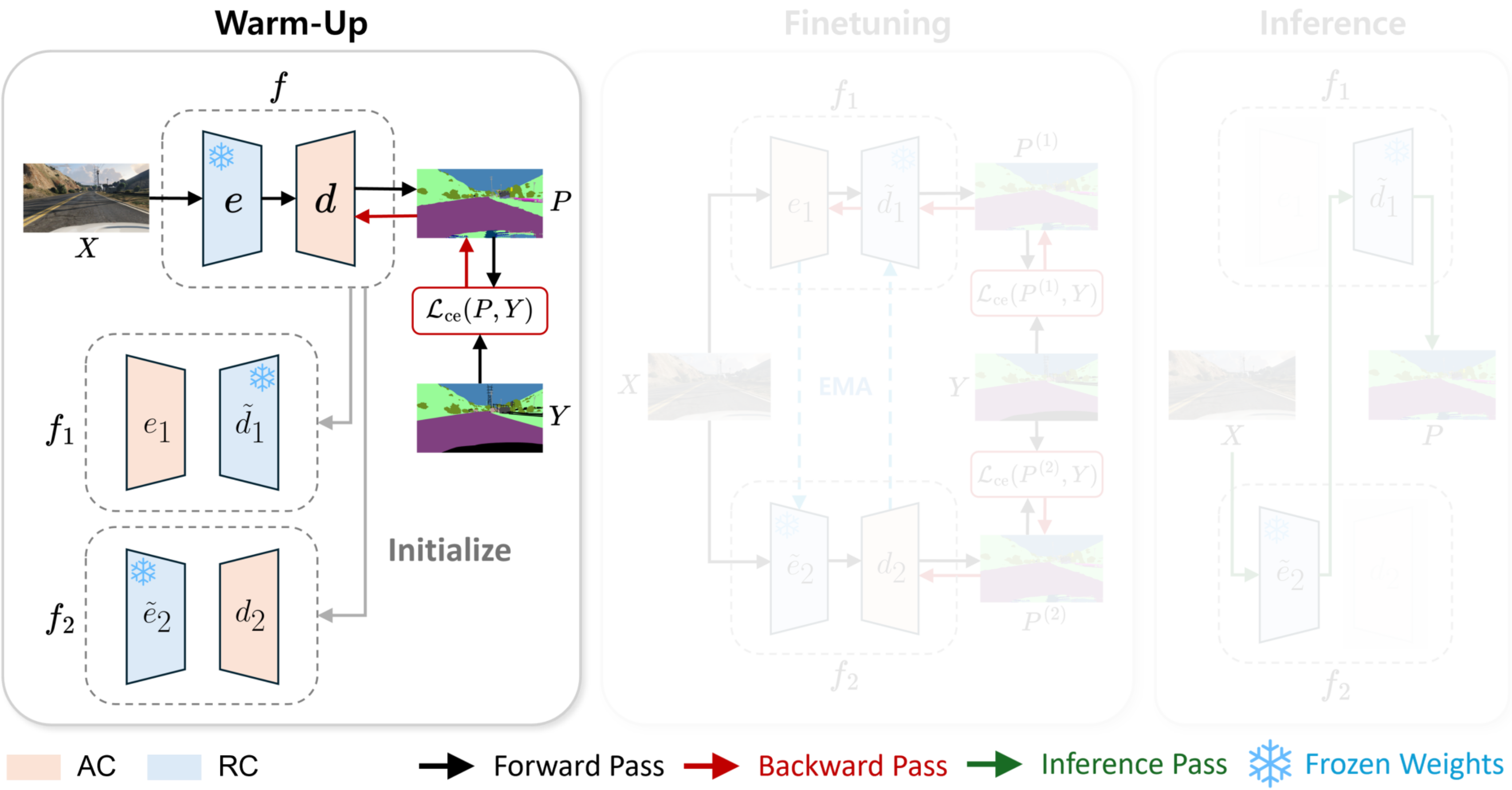


1. What should be used as the generalized counterpart?
   → ***EMA (Exponential Moving Average) versions of the encoder and decoder***
2. How can the alignment of the decoupled encoder and decoder be ensured?
   → ***Using the combination of the EMA versions as the final model***

Decoupled Finetuning (DeFT) Framework



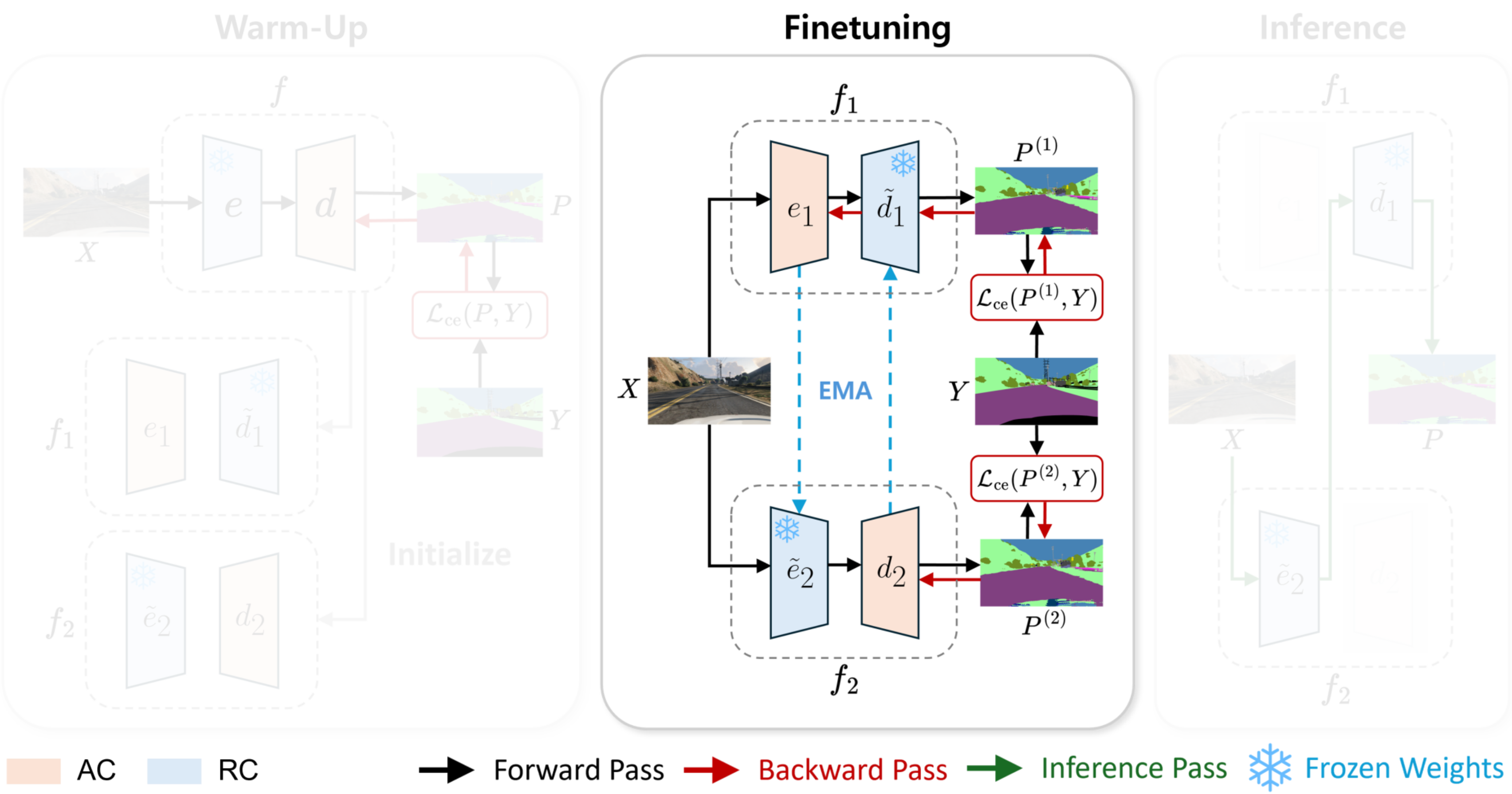
Decoupled Finetuning (DeFT) Framework



Stage 1. *Decoder Warm-Up**

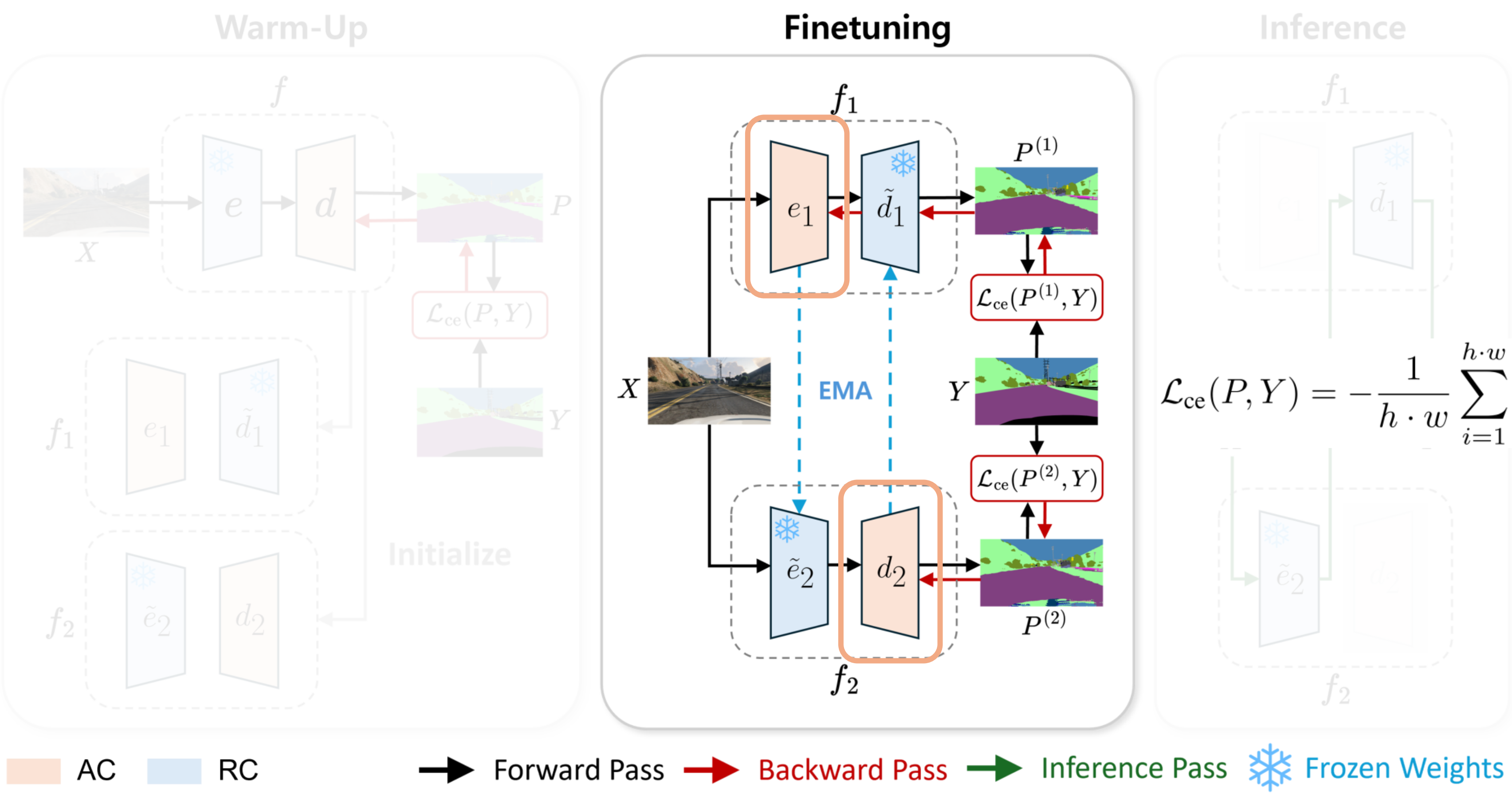
** Kumar et al., Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution. ICLR 2022*

Decoupled Finetuning (DeFT) Framework



Stage 2. *Decoupled Finetuning*

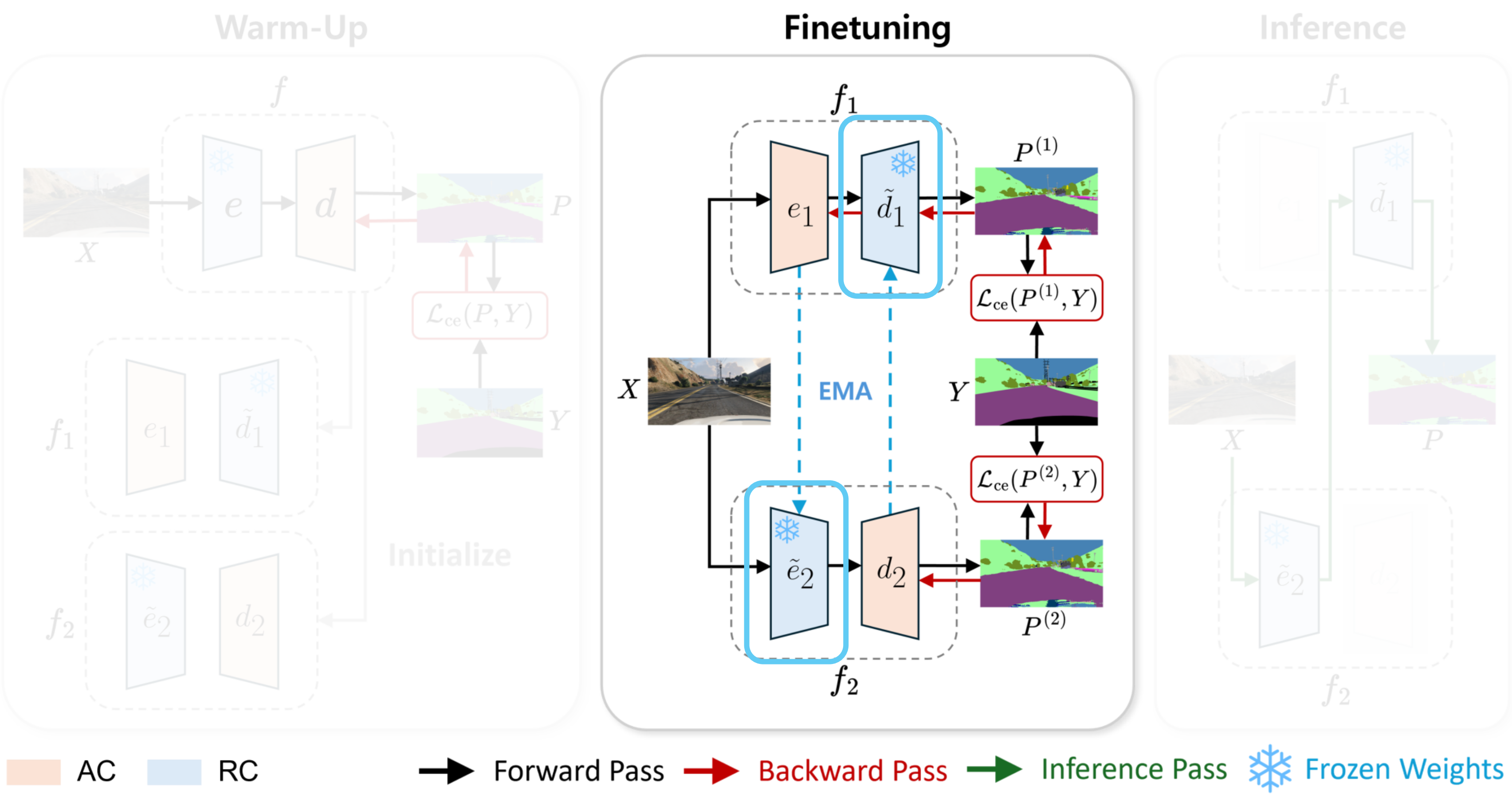
Decoupled Finetuning (DeFT) Framework



$$\mathcal{L}_{ce}(P, Y) = -\frac{1}{h \cdot w} \sum_{i=1}^{h \cdot w} Y_i^\top \cdot \log(P_i)$$

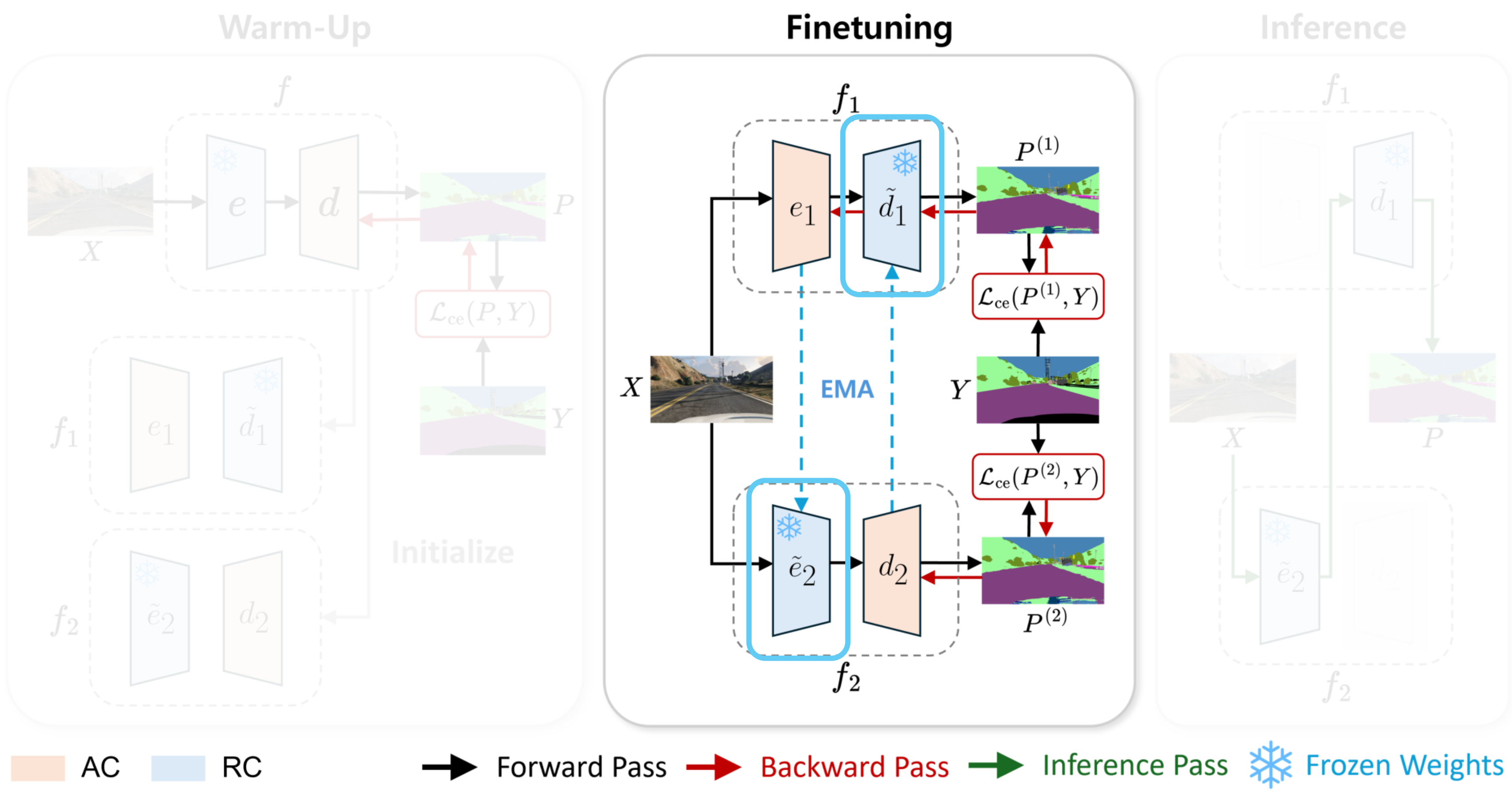
Stage 2-1. ***Adaptive Components (ACs)*** *are updated using gradients of the training loss*

Decoupled Finetuning (DeFT) Framework



Stage 2-2. ***Retentive Components (RCs)*** *are updated by the EMA of their counterpart ACs*

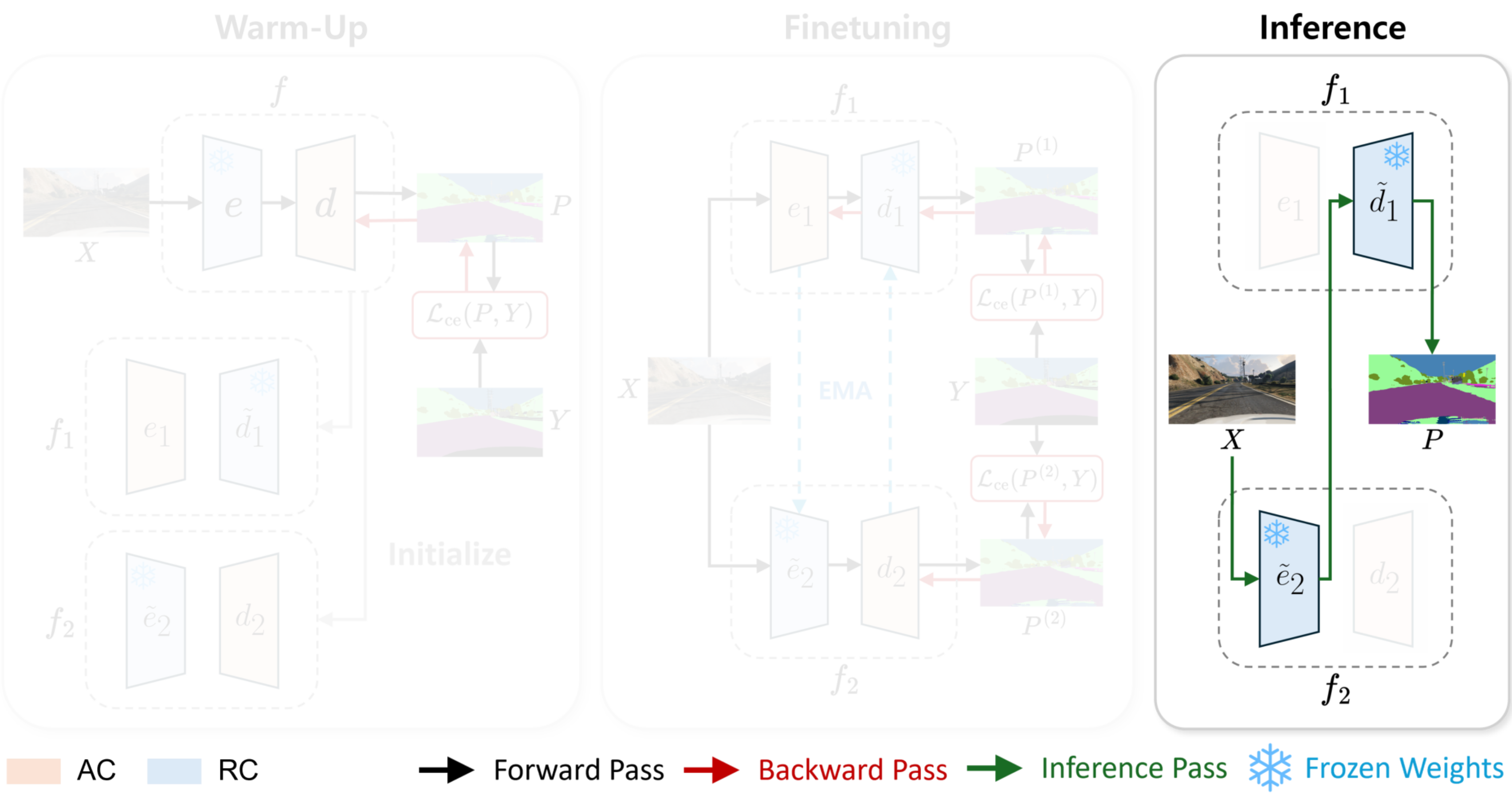
Decoupled Finetuning (DeFT) Framework



Stage 2-2. ***Retentive Components (RCs)*** *are updated by the EMA of their counterpart ACs*

$$\tilde{\theta}_{d_1}^{t+1} = \beta\tilde{\theta}_{d_1}^{t} + (1-\beta)\theta_{d_2}^{t}, \quad \tilde{\theta}_{e_2}^{t+1} = \beta\tilde{\theta}_{e_2}^{t} + (1-\beta)\theta_{e_1}^{t}$$

Decoupled Finetuning (DeFT) Framework



Inference. ***<u>Using the combination of the RCs as the final model</u>***

Why Does DeFT Work? (1)

(Joint Finetuning) A single optimization objective for all the parameters in the model

(DeFT) Two separate optimization objectives: <u>one for the encoder</u> and <u>the other for the decoder</u>

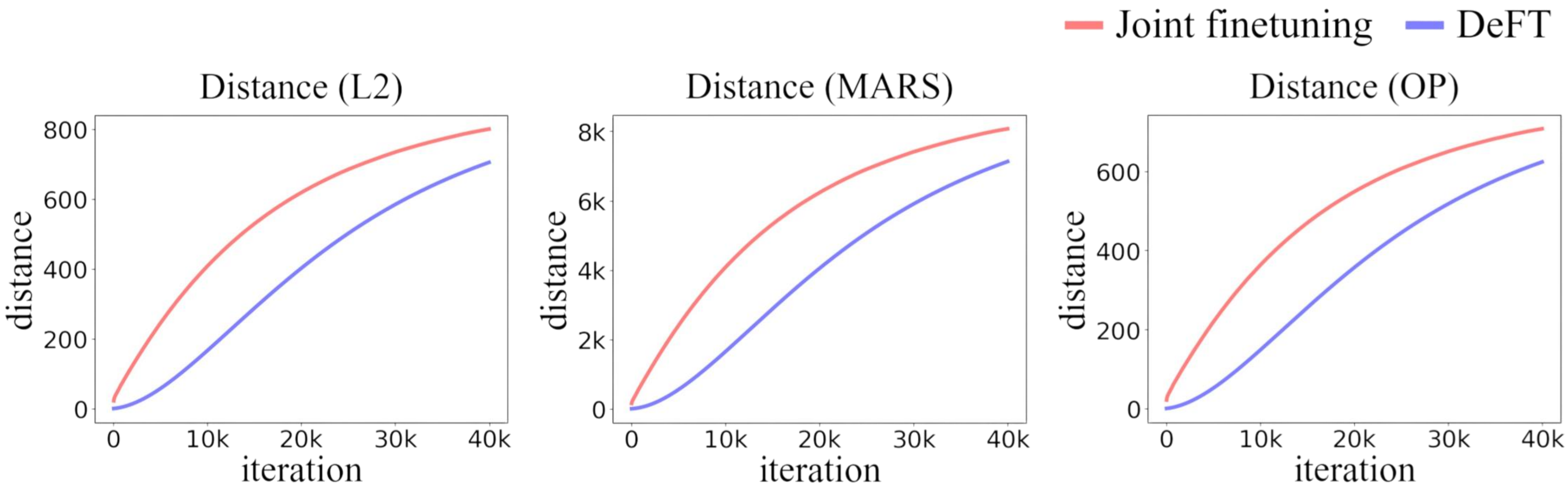
In **DeFT**, each trainable module is trained on a separate objective with ***fewer parameters***

→ ***Tighter generalization bound for each module[1] [2]***

[1] Du et al., How many samples are needed to estimate a convolutional neural network? NeurIPS 2018

[2] Long & Sedghi, Generalization bounds for deep convolutional neural networks. ICLR 2020

Why Does DeFT Work? (2)



Parameters with ***shorter distance from their initialization***

→ Tighter generalization bound for the final model[1] [2] [3]

[1] Nagarajan & Kolter, Generalization in deep networks: The role of distance from Initialization. 2019

[2] Long & Sedghi, Generalization bounds for deep convolutional neural networks. ICLR 2020

[3] Gouk et al., Distance-based regularisation of deep networks for fine-tuning. ICLR 2021

Experimental Results

Methods	ResNet-50				ResNet-101			
	C	B	M	Avg.	C	B	M	Avg.
Baseline	35.16	29.71	31.29	32.05	35.73	34.06	33.42	34.40
IBN-Net (Pan et al., 2018)	33.85	32.30	37.75	34.63	37.37	34.21	36.81	36.13
DRPC (Yue et al., 2019a)	37.42	32.14	34.12	34.56	42.53	38.72	38.05	39.77
ISW (Choi et al., 2021)	36.58	35.20	40.33	37.37	37.20	33.36	35.57	35.38
WildNet (Lee et al., 2022)	44.62	38.42	46.09	43.04	45.79	41.73	47.08	44.87
SAN-SAW (Peng et al., 2022)	39.75	37.34	41.86	39.65	45.33	41.18	40.77	42.43
DIRL (Xu et al., 2022)	41.04	39.15	41.60	40.60	-	-	-	-
SHADE (Zhao et al., 2022)	44.65	39.28	43.34	42.42	46.66	43.66	45.50	45.27
PASTA (Chattopadhyay et al., 2023)	44.12	40.19	47.11	43.81	45.33	42.32	48.60	45.42
TLDR (Kim et al., 2023b)	46.51	42.58	46.18	45.09	47.58	44.88	48.80	47.09
BlindNet (Ahn et al., 2024)	45.72	41.32	47.08	44.71	-	-	-	-
DeFT (Ours)	**50.06**	**43.17**	**50.51**	**47.91**	**52.14**	**45.16**	**53.15**	**50.15**

GTAV ➜ {**C**ityscapes, **B**DD100K, **M**apillary}

Methods	B	S	G	Avg.
Baseline	44.96	23.29	42.55	36.93
IBN-Net (Pan et al., 2018)	48.56	26.14	45.06	39.92
DRPC (Yue et al., 2019a)	49.86	26.58	45.62	40.69
ISW (Choi et al., 2021)	50.74	26.20	45.00	40.64
WildNet (Lee et al., 2022)	50.94	27.95	47.01	41.97
SAN-SAW (Peng et al., 2022)	52.95	28.32	47.28	42.85
DIRL (Xu et al., 2022)	51.80	26.50	46.52	41.61
SHADE (Zhao et al., 2022)	50.95	27.62	48.61	42.39
BlindNet (Ahn et al., 2024)	51.84	28.51	47.97	42.77
DeFT (Ours)	**53.12**	**28.87**	**48.72**	**43.57**

Cityscapes ➜ {**B**DD100K, **S**YNTHIA, **G**TAV}

Ablation Study (1)

w/o Aux.	Aug.	Warm-Up	DeFT	Cityscapes	BDD100K	Mapillary	Avg.
				35.16	29.71	31.29	32.05
✓				36.58	34.49	39.08	36.72
✓	✓			40.77	37.87	43.39	40.66
✓	✓	✓		42.32	40.33	44.88	42.51
✓	✓	✓	✓	**50.06**	**43.17**	**50.51**	**47.91**

The impact of individual component

Ablation Study (2)

Finetuning strategy	Cityscapes	BDD100K	Mapillary	Avg.
Joint finetuning	42.32	40.33	44.88	42.51
Joint finetuning + EMA	48.30	42.29	49.02	46.54
DeFT	**50.06**	**43.17**	**50.51**	**47.91**

The impact of the decoupled finetuning strategy

Ablation Study (3)

ID	e_1 (AC)	$\tilde{e}_2$ (RC)	$\tilde{d}_1$ (RC)	d_2 (AC)	Cityscapes	BDD100K	Mapillary	Avg.
I	✓			✓	39.30	37.41	43.14	39.95
II	✓		✓		43.15	39.82	45.55	42.84
III		✓		✓	<u>47.29</u>	<u>41.84</u>	<u>49.33</u>	<u>46.15</u>
IV		✓	✓		**50.06**	**43.17**	**50.51**	**47.91**

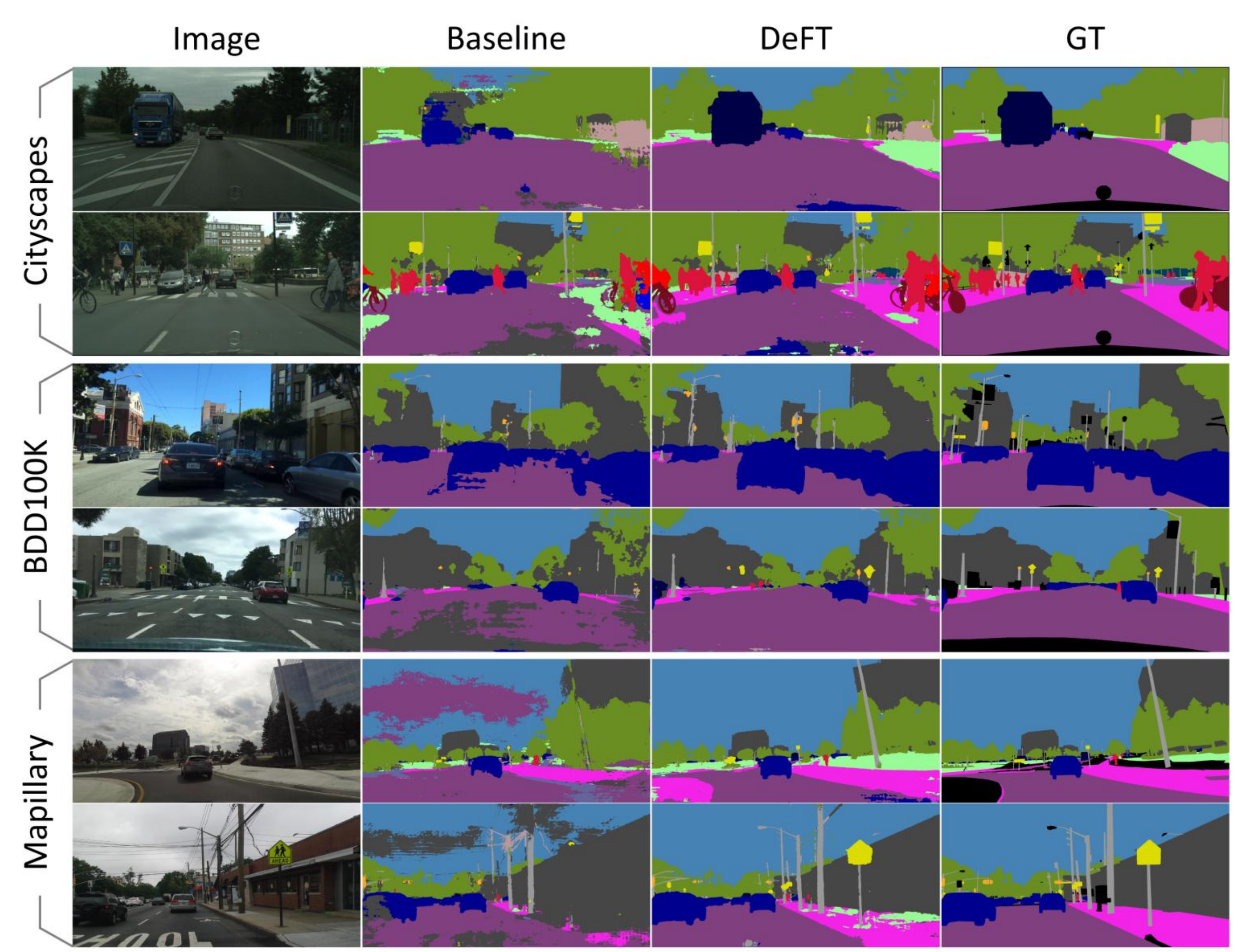
The impact of final model configuration

Ablation Study (4)

EMA update ratio (β)	Cityscapes	BDD100K	Mapillary	Avg.
0.99	44.37	40.79	46.81	43.99
0.999	46.19	42.14	48.81	45.71
0.9999	**50.06**	**43.17**	**50.51**	**47.91**

The impact of the EMA update ratio β

Qualitative Results



Conclusion

- We have demonstrated **the detrimental effects of jointly finetuning** the encoder and decoder.
- We introduce **DeFT**, a novel and effective training framework that **decouples the finetuning of the encoder and decoder**.
- Building a **more concrete theoretical foundation** and exploring a **better alternative configurations for the RCs** will be promising future research directions.