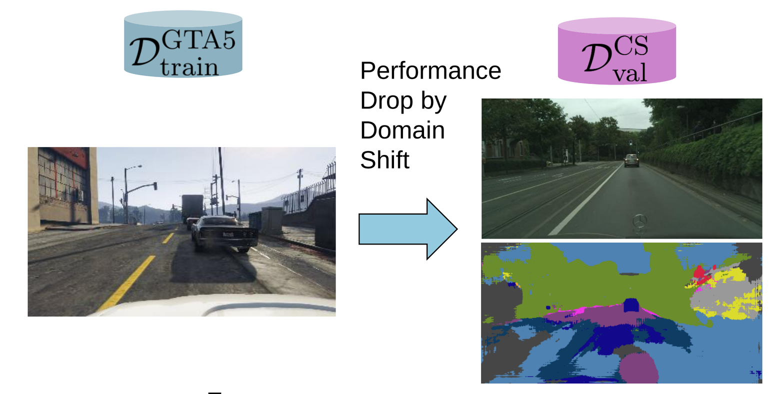
Strong but Simple: A Baseline for Domain Generalized Dense Perception by CLIP-based Transfer Learning

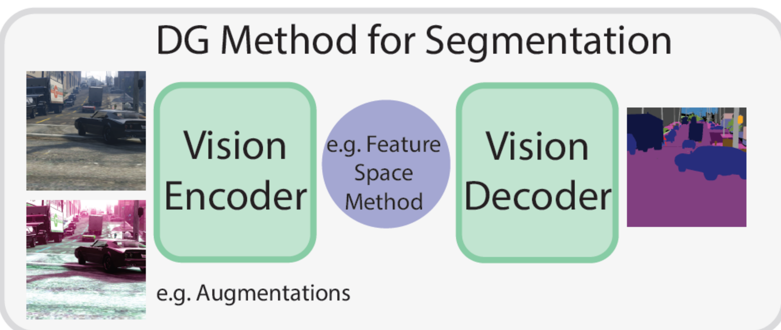
Christoph Hümmer*, Manuel Schwonberg*, Liangwei Zhou*, Hu Cao, Alois Knoll and Hanno Gottschalk

ACCV HANOI VIETNAM 2024 DEC 8-12

1 Contribution & Proposed Method

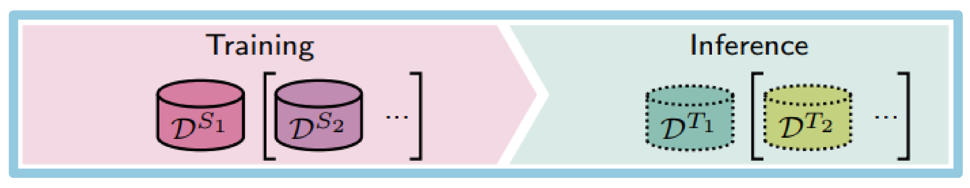


Previous



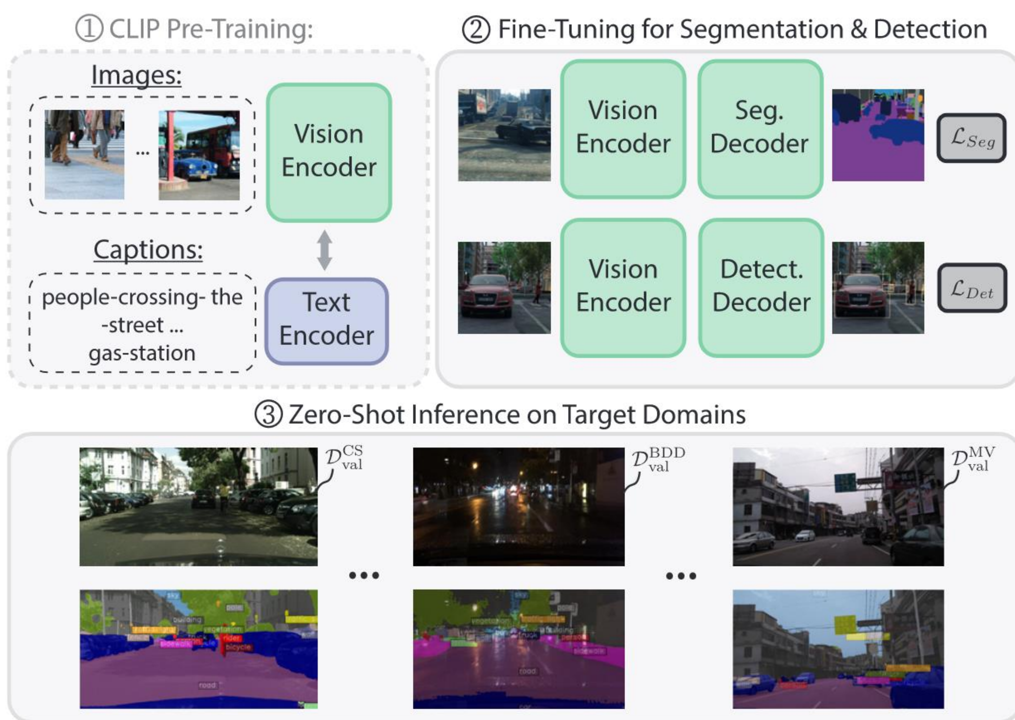
→ PREVIOUS approaches require complex methods or modules

→ Ours only relies on fine-tuning a strong vision-language pre-trained backbone



Domain Generalization (DG)
+ No target data/knowledge required
+ Generalization across multiple unknown domains

Ours



Our **key contributions** are:

- By **simple fine-tuning WITHOUT any additional modules or methods for segmentation AND detection** our approach shows similar or stronger performance than previous works.
- **New SOTA performance on Cityscapes, ACDC and synthetic-to-real** benchmarks and proposing new real-to-real and synthetic-to-real evaluation scheme

2 Comparison of Pre-Trainings

$\mathcal{D}^S$	Pre-Training Method	Data	Sup.	Self-Sup.	$\mathcal{D}^{CS}_{val}$	$\mathcal{D}^{BDD}_{val}$	$\mathcal{D}^{MV}_{val}$	$\mathcal{D}^{ACDC}_{val}$	DG mean
GTA5	SegFormer [105]	ImgNet-1K	✓	✗	46.6	45.6	50.1	36.4	44.7
	Supervised*	ImgNet-21K	✓	✗	49.3	47.0	52.2	43.5	48.0
	MoCov3* [13]	ImgNet-1K	✗	✓	49.7	46.2	52.4	39.1	46.9
	DeiT3* [94]	ImgNet-21K	✓	✓	53.7	52.6	59.3	49.0	53.7
	SAM* [53]	SA-1B	✓	✗	53.2	50.3	58.8	45.5	52.0
	LDM [77]	LAION-5B	✓	✗	49.2	-	-	-	-
	CLIP 🔒° [73]	WIT	✓	✗	53.2	49.8	57.1	45.0	51.2
	CLIP ° [73]	WIT	✓	✗	55.6	52.5	59.9	51.5	54.9
	EVA-02-CLIP 🔒° [90]	Merged-2B	✓	✗	55.2	51.3	57.4	47.3	52.8
	EVA-02-CLIP° [90]	Merged-2B	✓	✗	**65.3**	**58.3**	**66.0**	**62.6**	**63.1**
SYNTHIA	SegFormer [105]	ImgNet-1K	✓	✗	41.1	36.2	42.4	32.6	38.1
	Supervised*	ImgNet-21K	✓	✗	44.3	37.1	43.1	34.8	39.8
	MoCov3* [13]	ImgNet-1K	✗	✓	40.2	35.4	41.5	31.7	37.2
	DeiT3* [94]	ImgNet-21K	✓	✓	47.8	39.1	45.4	34.7	41.8
	SAM* [53]	SA-1B	✓	✗	51.6	40.4	50.1	40.1	45.6
	CLIP 🔒° [73]	WIT	✓	✗	46.1	41.8	45.8	35.1	42.2
	CLIP ° [73]	WIT	✓	✗	51.1	44.7	50.6	40.7	46.8
	EVA-02-CLIP 🔒° [90]	Merged-2B	✓	✗	48.3	42.6	46.4	37.1	43.6
	EVA-02-CLIP ° [90]	Merged-2B	✓	✗	**56.8**	**51.9**	**55.1**	**48.5**	**53.1**

+9.4% +7.5%

→ Vision-Language pre-training performs significantly better than vision-only and benefits from large-scale text-image datasets

3 DG Benchmarks

	DG Method	Enc. Params	$\mathcal{D}^{CS}_{val}$	$\mathcal{D}^{BDD}_{val}$	$\mathcal{D}^{MV}_{val}$	DG mean
Vision DG	Baseline	81.4M	46.6	45.6	50.1	47.4
	ReVT	81.4M	50.0	48.0	52.8	50.3
	SHADE	81.4M	53.3	48.2	55.0	52.2
	IBAFormer	81.4M	56.3	49.8	58.3	54.8
	CMFormer	197M	55.3	49.9	60.1	55.1
	HRDA	81.4M	57.4	49.1	61.2	55.9
Vision-Language DG	DGinStyle	81.4M	58.6	52.3	62.5	57.8
	DIDEX	81.4M	62.0	54.3	62.0	59.7
	PromptFormer	-	52.0	-	-	-
	CLOUDS	198M	60.2	57.4	**67.0**	61.5
	Rein	307M	**65.3**	**60.5**	64.9	**63.6**
	VLTSeg	304M	**65.3**	58.3	66.0	63.2

SOTA-Level

→Competitive to SOTA and significantly better than other, more complex works

Method	Encoder	Init	Day Clear	Night Clear	Dusk Rainy	Night Rainy	Day Foggy	Average
S-DGOD [103]	R-101	IN1k	56.1	36.6	28.2	16.6	33.5	28.7
G-NAS [104]	R-101	IN1k	58.4	**45.0**	35.1	17.4	36.4	33.5
PDDOC [58]	R-101	IN1k	53.6	38.5	33.7	19.2	39.1	32.6
CLIP The GAP [97]	R-101	CLIP	51.3	36.9	32.3	18.7	38.5	31.6
PODA [26]	R-101	CLIP	-	43.4	40.2	20.5	**44.4**	37.1
VLTDet	R-101	CLIP	**60.5**	44.6	38.4	22.1	42.3	36.9
VLTDet	ViT-L-14	EVA02-CLIP	56.6	44.4	**43.6**	**26.6**	41.8	**39.1**

+2.1% +2.0%

New SOTA

Source	Method	Seg. mIoU $\mathcal{D}^{CS}_{val}$	$\mathcal{D}^{BDD}_{val}$	$\mathcal{D}^{MV}_{val}$	$\mathcal{D}^{ACDC}_{val}$	rPD	Method	Det. mAP_{50} $\mathcal{D}^{CS}_{val}$	$\mathcal{D}^{BDD}_{val}$	$\mathcal{D}^{ACDC}_{val}$	rPD
$\mathcal{D}^{CS}_{train}$	SegFormer	80.8	58.0	69.1	58.6	76.6	IN21k	52.8	30.8	35.6	62.9
	SAM+M2F	82.8	57.1	71.0	60.5	75.9	SAM	**61.2**	37.2	**45.1**	67.2
	VLTSeg	**85.0**	**66.0**	**77.2**	**73.3**	**84.9**	VLTDet	59.7	**39.6**	44.3	**70.2**
$\mathcal{D}^{BDD}_{train}$	SegFormer	62.4	64.4	63.0	54.2	93.0	IN21k	45.2	51.2	35.7	79.0
	SAM+M2F	68.8	69.3	69.1	57.9	94.2	SAM	**47.5**	**53.8**	**37.1**	78.6
	VLTSeg	**78.9**	**72.5**	**75.9**	**71.3**	**104.0**	VLTDet	46.6	49.2	34.5	**82.4**
$\mathcal{D}^{ACDC}_{train}$	SegFormer	66.3	54.1	64.9	75.0	82.4	IN21k	22.4	22.5	32.8	68.4
	SAM+M2F	68.9	57.0	68.1	77.4	83.5	SAM	33.9	25.3	47.8	61.9
	VLTSeg	**78.8**	**65.0**	**74.1**	**80.0**	**90.8**	VLTDet	**44.9**	**31.0**	**48.3**	**78.6**
$\mathcal{D}^{MV}_{train}$	SegFormer	76.1	63.0	78.9	65.7	86.5	-	-	-	-	-
	SAM+M2F	76.4	62.8	78.9	63.4	85.6	-	-	-	-	-
	VLTSeg	**81.6**	**69.2**	**83.8**	**74.2**	**89.5**	-	-	-	-	-

+8.3% +10.4%

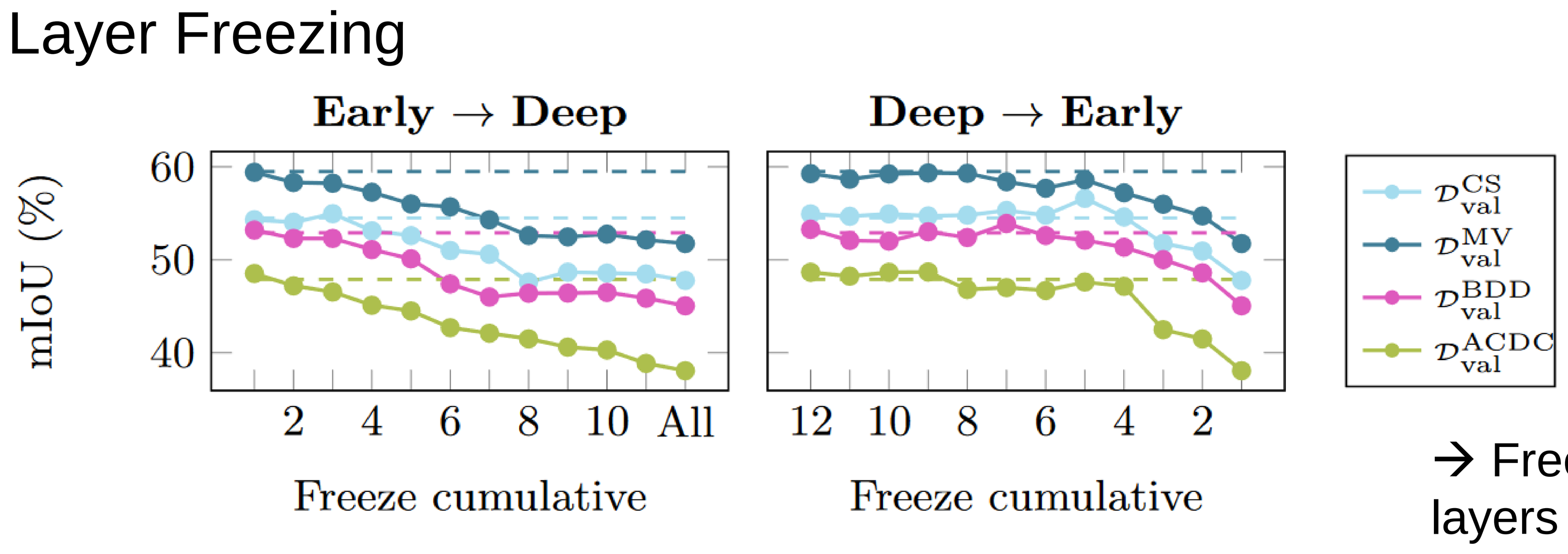
→ On par with SOTA for ResNet-101 for object detection, SOTA with EVA-backbone

→ Signifcantly enhanced real-to-real generalization

→Enhanced in-domain performance

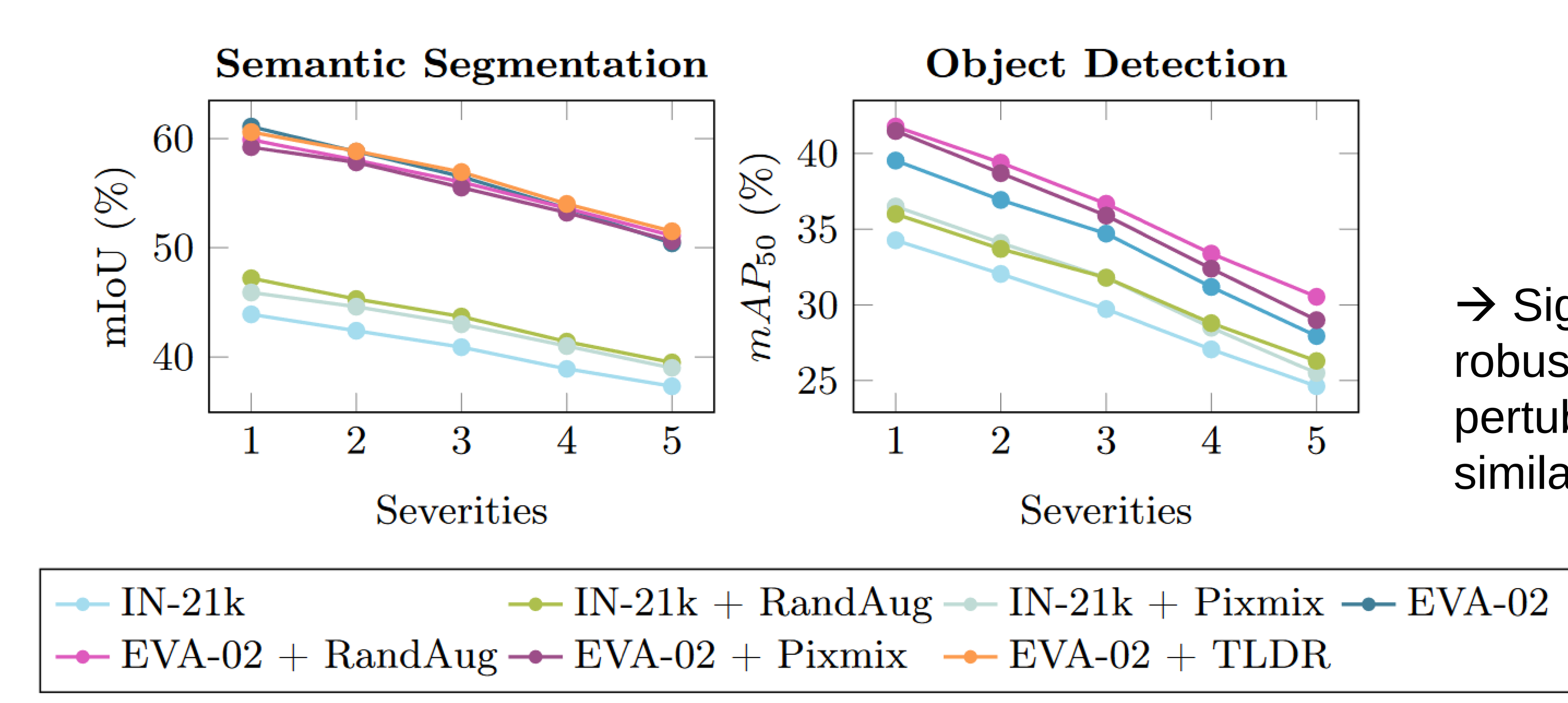
4 Analysis

Layer Freezing



→ Freezing earlier layers decreases performance

Pertubations



→ Significantly more robust against pertubations but similar trend

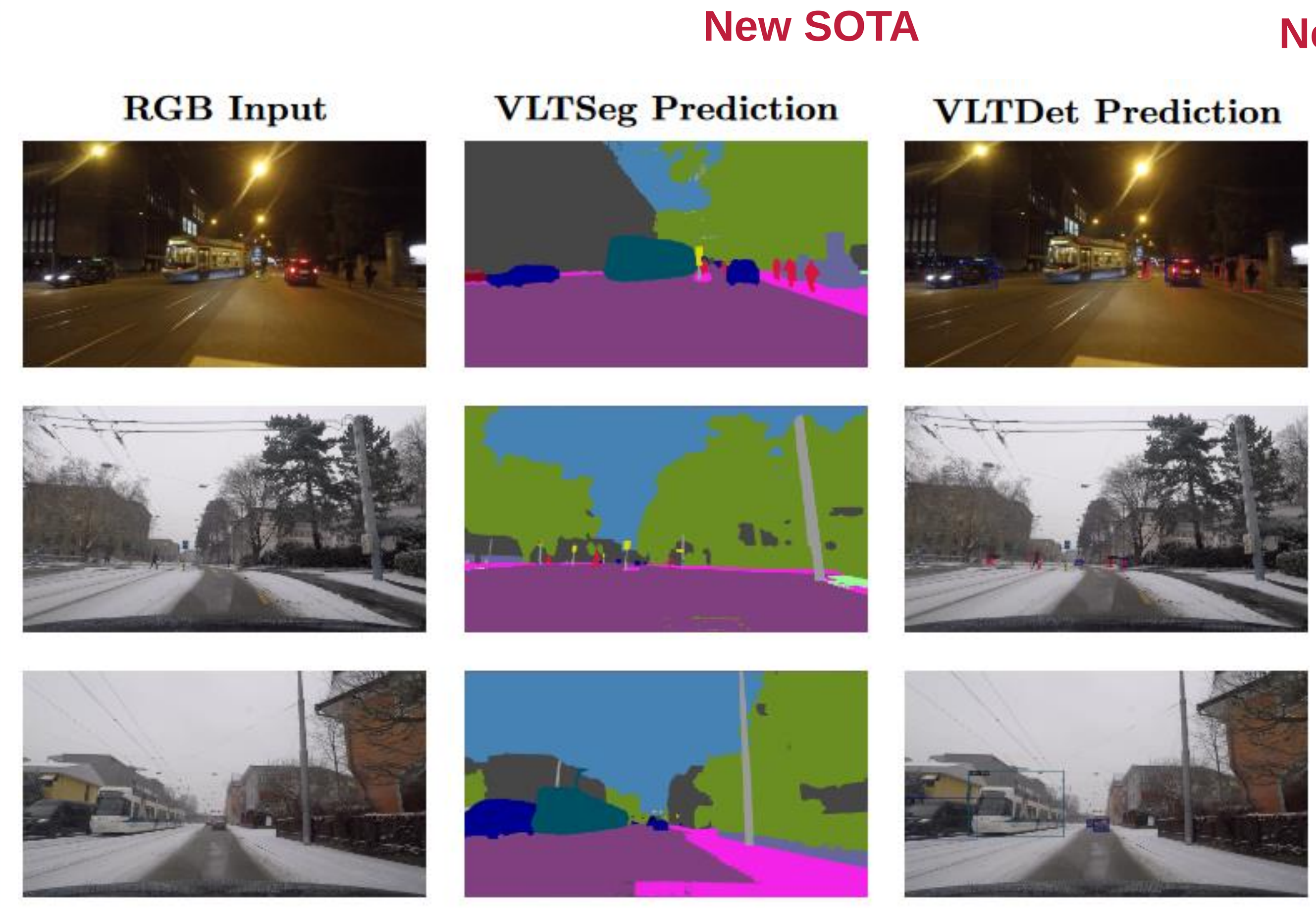
5 Real-to-Real DG

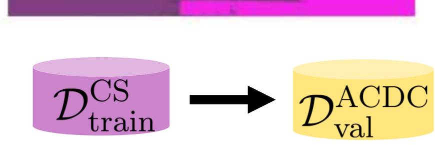
Cityscapes ($\mathcal{D}^{CS}_{train}$) → Cityscapes ($\mathcal{D}^{CS}_{test}$) Method	Rank	Params	Iter.	mIoU in %
ViT-Adapter-L [14]	5	571M	80k	85.2
InverseForm [7]	3	-	-	85.6
HS3 [6]	4	-	-	85.8
InternImage [99]	2	1.2B	80k	86.1
VLTSeg	1	304M	40k	**86.4**

Cityscapes ($\mathcal{D}^{CS}_{train}$) → ACDC ($\mathcal{D}^{ACDC}_{test}$) Method	UDA/DG	mIoU in %
HRDA [41]	UDA	67.96
CISS [80]	UDA	69.55
PromptFormer [32]	DG	62.0
Rein [102]	DG	77.56
VLTSeg (1024^2)	DG	**77.91**

+0.3% +0.35%

New SOTA New SOTA





Paper

Project Page