Doubly Contrastive End-to-End Semantic Segmentation for Autonomous Driving under Adverse Weather

Jongoh Jeong and Jong-Hwan Kim

Korea Advanced Institute of Science and Technology (KAIST)

BMVC 2022

INTRODUCTION

Motivation

- Intelligent driving systems require safe and accurate perception of the surroundings in dynamically changing environments, most heavyweight models that focus primarily on the performance are not suitable for practical real-time deployment.
- Most semantic segmentation applications under adverse driving conditions, or “unusual road or traffic conditions that were not known" as defined by the US Federal Motor Carrier Safety Administration [1] (e.g., fog, nighttime, rain, snow), has yet to be explored further [2].
- 1.3 million death toll of road traffic crashes every year, and the risk of accidents in rainy weather, for example, is 70% higher than in normal conditions [3, 4]

Contributions

- We propose an end-to-end doubly (image- and pixel-levels) contrastive learning strategy for a lightweight semantic segmentation model to eliminate the pre-training stage in the conventional contrastive learning approach without requiring a large training batch size or a memory bank.
- Our training method achieves 1.34%p increase in mIoU measure from the baseline focal loss-only objective with the SwiftNet architecture (ResNet-18 backbone), running inference at up to 66.7 FPS in 2048×1024 resolution on a single Nvidia RTX 3080 Mobile GPU.
- We verify that replacing image-level supervision with self-supervision in our supervised contrastive objective achieves comparable performance when pre-trained with clear weather images.

PRELIMINARIES

- Self-supervised Contrast
  - For a set of N randomly sampled image-label pairs $\{\boldsymbol{X_k}, \boldsymbol{y_k}\}_{k=1,\dots N}$ and a corresponding multi-viewed set of augmented samples from the same sources $\{\widetilde{\boldsymbol{X_k}}, \widetilde{\boldsymbol{y_k}}\}_{k=1,\dots 2N}$, compute a similarity score

$$\mathcal{L}_{self} = \sum_{i\in I} \mathcal{L}_{self}^{(i)} = -\sum_{i\in I} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_{j(i)}/\tau)}{\sum_{a\in A(i)} \exp(\mathbf{z}_i \cdot \mathbf{z}_a/\tau)}$$

- Supervised Contrast
  - Image-level
    - More than one sample belonging to each image class label

$$\mathcal{L}_{image} = \sum_{i\in I} \mathcal{L}_{image}^{(i)} = -\sum_{i\in I} \frac{1}{|P(i)|} \sum_{p\in P(i)} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_p/\tau)}{\sum_{a\in A(i)} \exp(\mathbf{z}_i \cdot \mathbf{z}_a/\tau)}$$

  - Pixel-level
    - Similar to image-level, but applied to pixel-wise features

$$\mathcal{L}_{pixel} = -\frac{1}{|P(i)|} \sum_{\mathbf{i}_p\in P(i)} \log \frac{\exp(\mathbf{i} \cdot \mathbf{i}_p/\tau)}{\sum_{\mathbf{i}_a\in A(i)} \exp(\mathbf{i} \cdot \mathbf{i}_a/\tau)} \qquad \forall i \in \{1,\dots,H\times W\},$$

METHODOLOGY

Doubly Contrastive Supervised Semantic Segmentation

- Base model: multi-scale pyramidal encoder followed by up-sampling blocks (SwiftNet)
- Training objective: Segmentation loss + Contrastive Losses(*see Preliminaries)
  - Applies contrastive objectives in image- and pixel-levels with supervision (ground truth weather class and pixel-level semantic labels, respectively)

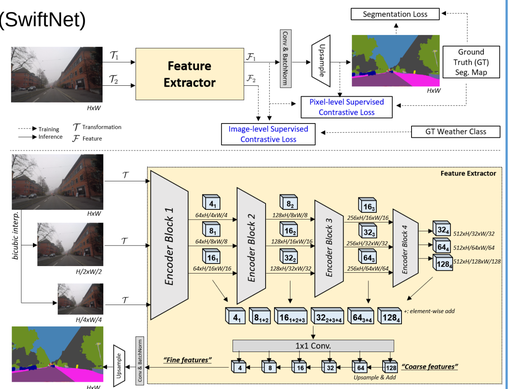
$$\mathcal{L} = \lambda_c \cdot (\mathcal{L}_{\text{image}} + \mathcal{L}_{\text{pixel}}) + \lambda_s \cdot \mathcal{L}_{seg},$$

where

$$\mathcal{L}_{seg}\left(\phi(p), \hat{\phi}(p)\right) = -\delta(p) e^{\gamma(1-P_t)} log(P_t),$$

$$\delta(p) = \underbrace{log\left(1+\varepsilon+\frac{freq_c(p)}{N_p}\right)^{-1}}_{\text{Class balancing}} \cdot \underbrace{exp\left(-\frac{d_{EDT}(p)}{2\sigma_{EDT}}\right)}_{\text{EDT}},$$

$$d_{EDT}(p) = \sum_C \min_{q\in C} \|p-q\|_2 \qquad \forall p \in G_{H\times W},$$



EXPERIMENTAL RESULTS

Datasets

- Cityscapes (pre-training)
- Adverse Conditions Dataset with Correspondences (ACDC) for training & evaluation

Table 1: Dataset Description

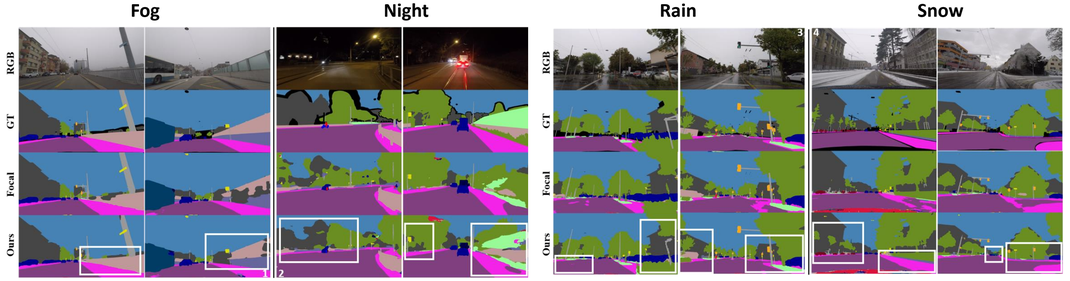
Dataset	Input Modality	Resolution	Anno. (# Classes)	Weather Condition	Train	Val	Test
Cityscapes [10]	Stereo RGB	2048×1024	Fine (19)	Clear/Daytime	2,975	500	1,525
ACDC [35]	Monocular RGB	1920×1080	Fine (19)	All	1,600	406	2,000
				Fog	400	100	500
				Nighttime	400	106	500
				Rain	400	100	500
				Snow	400	100	500

Table 2: Semantic segmentation performance by adverse weather conditions using SwiftNet. Bb denotes backbone. The best results in **boldface** and the second best in underline.

Bb	Exp.	Loss (*: single contrast, †: double contrasts)	mIoU (%)	Fog	Nighttime	Rain	Snow
ResNet-18	(a)	Cross Entropy	62.63	68.95	45.71	63.32	65.43
	(b)	Focal (baseline)	64.04	71.59	47.84	63.90	65.57
	(c)	+Pixel-level Supervised Contrast only*	63.79	69.76	47.26	63.05	68.34
	(d)	+Self-supervised Contrast only*	63.91	71.83	48.02	62.24	68.35
	(e)	+Image-level Supervised Contrast only*	63.62	69.66	47.65	63.28	67.10
	(f)	+Self-supervised and Pixel-level Supervised Contrasts†	65.07	**72.45**	**48.57**	63.95	68.31
	(g)	+Image- and Pixel-level Supervised Contrasts† *(Ours)*	**65.38**	67.94	48.56	**65.38**	**68.64**
ResNet-34	(a)	Cross Entropy	65.02	73.85	47.11	65.20	66.94
	(b)	Focal (baseline)	67.00	74.44	49.38	67.29	70.64
	(c)	+Pixel-level Supervised Contrast only*	66.60	72.49	49.84	65.55	69.06
	(d)	+Self-supervised Contrast only*	68.09	**76.99**	50.14	66.45	69.56
	(e)	+Image-level Supervised Contrast only*	68.07	75.12	50.76	66.78	69.33
	(f)	+Self-supervised and Pixel-level Supervised Contrasts†	67.02	72.06	50.12	65.86	70.83
	(g)	+Image- and Pixel-level Supervised Contrasts† *(Ours)*	**68.33**	75.19	**51.21**	**67.37**	**71.19**
Bb	**Exp.**	**Loss** (w/ pre-training with clear weather)	**mIoU (%)**	**Fog**	**Nighttime**	**Rain**	**Snow**
ResNet-18	-	Focal (*Cityscapes*)	73.16	N/A (Clear weather only)			
	(a)	Cross Entropy	64.40	71.88	47.09	65.59	66.66
	(b)	Focal (baseline)	65.49	73.43	47.67	64.77	68.35
	(f)	+Self-supervised and Pixel-level Supervised Contrasts†	**66.97**	74.05	**49.10**	**67.85**	**69.69**
	(g)	+Image- and Pixel-level Supervised Contrasts (*Ours*)†	66.24	**75.38**	48.82	65.79	68.42
ResNet-34	-	Focal (*Cityscapes*)	73.80	N/A (Clear weather only)			
	(a)	Cross Entropy	68.38	75.99	49.70	67.92	71.49
	(b)	Focal (baseline)	69.46	**76.94**	50.28	69.78	71.69
	(f)	+Self-supervised and Pixel-level Supervised Contrasts†	70.06	76.11	50.47	**70.68**	**72.89**
	(g)	+Image- and Pixel-level Supervised Contrasts (*Ours*)†	**70.13**	76.31	**53.59**	70.50	71.89

Results

Our approach achieves 1.34%p and 1.33%p increases in mIoU using the ResNet-18 and 34, respectively, from the baseline. In addition, we found that our method is more advantageous in more adverse conditions like nighttime, rain and snow than relatively easier condition like fog. We assume this is due to the difficulties where the visibility is limited by the low illumination, rain streaks, rain/snow-covered roads, and highly dense fog. Our approach corrects false positive predictions observed in the baseline results, and results in more consistent predictions of objects of relatively large sizes.



ABLATION STUDY

For ablation study, we compared our approach on ENet and DeepLabV3+ model architectures, which are based on single-scale and Atrous Spatial Pyramid Pooling (ASPP) encoders, respectively. We chose these models with a lighter backbone due to the real-time and lightweight constraints. From the table, we can see that application to ENet shows the self-supervised replacement in our approach performs the best, and DeepLabV3+ did not show much improvement. We posit that different encoder structures results in more or less discriminative features and our approach works best on multi-scale pyramidal type of encoder. DeepLabV3+'s ASPP, for example, simply applies different kernel-sized convolution operations on one set of features, while SwiftNet extracts features in multiple scales sequentially from the start.

Table 3: Ablation study: semantic segmentation performance with different models and *coarse* features (2048×1024 resolution). *Coarse* features are marked with †; otherwise *fine*.

Model (Encoder type)	Exp.	mIoU (%)	Fog	Nighttime	Rain	Snow	GFLOPs	Params (M)	RTX 3080 Mobile Time (ms)	FPS
ENet [29] (Single-scale)	(a)	45.22	48.10	36.31	44.46	47.53	1.40	0.35	31	32.3
	(b)	50.45	55.14	38.70	49.44	**53.44**				
	(f)	**50.78**	**55.91**	**38.96**	**50.88**	52.39				
	(g)	49.32	53.64	37.85	49.86	51.17				
SwiftNet (ResNet-18) [28] (Multi-scale pyramidal)	(a)	62.63	68.95	45.71	63.32	65.43	8.04	12.04	15	66.7
	(b)	64.04	71.59	47.84	63.90	65.57				
	(f)	65.07	**72.45**	**48.57**	63.95	68.31				
	(g)	**65.38**	67.94	48.56	**65.38**	**68.64**				
	(f)†	60.62	65.84	45.35	61.20	64.79				
	(g)†	61.66	67.86	46.07	62.10	64.84				
SwiftNet (ResNet-34) [28] (Multi-scale pyramidal)	(a)	65.02	73.85	47.11	65.20	66.94	14.40	22.15	26	38.5
	(b)	67.00	74.44	49.38	67.29	70.64				
	(f)	67.02	72.06	50.12	65.86	70.83				
	(g)	**68.33**	**75.19**	**51.21**	**67.37**	**71.19**				
DeepLabV3+ (ResNet-50) [7] (ASPP)	(a)	69.69	73.60	51.45	72.18	72.29	29.88	39.76	14	71.4
	(b)	**70.07**	**75.18**	**52.49**	**73.22**	71.50				
	(f)	69.22	73.95	50.54	70.42	**73.24**				
	(g)	69.04	74.54	51.20	70.70	71.83				

CONCLUSIONS

We proposed an end-to-end doubly contrastive learning approach to semantic segmentation for self-driving under adverse weather, exploiting image-level labels to semantically correlate RGB images taken under various weather conditions and pixel-level labels to obtain more semantically meaningful representations. In our method, the two supervised contrasts complement each other to effectively improve the performance of a lightweight model, without a need for pre-training or a memory bank to associate images across various weather conditions for global consistency.

REFERENCES

1. Federal Motor Carrier Safety Administration. code of federal regulations. Accessed May 23, 2022 [Online].
2. Christos Sakaridis, Dengxin Dai, and Luc Van Gool. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. ICCV, October 2021.
3. Jean Andrey and Sam Yagar. A temporal analysis of rain-related crash risk. Accident Analysis & Prevention, 25(4):465–472, 1993. ISSN 0001-4575. doi: https:// doi.org/10.1016/0001-4575(93)90076-9.
4. World Health Organization. Road traffic injuries, 2021.

ACKNOWLEDGMENTS

This work was supported by the Institute for Information & Communications Technology Promotion (IITP) under Grant 2020-0-00440 through the Korean Government (Ministry of Science and ICT (MSIT); Development of Artificial Intelligence Technology that Continuously Improves Itself as the Situation Changes in the Real World).