Special Paper
방송공학회논문지 제29권 제7호, 2024년 12월 (JBE Vol. 29, No. 7, December 2024)
https://doi.org/10.5909/JBE.2024.29.7.1136
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

A Study on the Vulnerability of Semantic Segmentation Model to Data Transformation

Chaewon Moon[a)], Dong-hwi Kim[a)], Dabin Kang[a)], and Sang-hyo Park[a)‡]

Abstract

With the advancement of autonomous driving technology, the importance of semantic segmentation has markedly increased, while the amount of datasets needed for training has been limited. Accordingly, there has been a growing effort to increase datasets using data augmentation techniques to train semantic segmentation models. However, the distributional gap between augmented and real data can lead to performance limitations when models trained on real data are applied to augmented data. Therefore, this paper constructs new datasets by applying proposed data transformations on real-world datasets. Additionally, we evaluate the impact of these transformations on semantic segmentation models trained on real datasets. Results show that semantic segmentation models are vulnerable to distortions in color information and object characteristics in transformed datasets. Furthermore, the vision transformer based model is less sensitive to distribution changes and shows greater segmentation performance compared to fully convolutional network based models.

Keyword : Semantic Segmentation, Autonomous Driving, Data Transformation, ViT, CNN

Ⅰ. Introduction

Semantic segmentation aims to predict the semantic category of each pixel in an input image and segment the image accordingly. This technique is widely utilized across various fields, including robotics, medical applications, and satellite image analysis, with particular prominence in autonomous driving[1]. In particular, autonomous driving is a rapidly growing field, driven by advancements in deep learning, where accurately understanding road scenes is a crucial challenge. Moreover, autonomous driving systems are likely to face a wide variety of situations on the road, highlighting the importance of large-scale road datasets that capture diverse driving scenarios[2]. To build such datasets, data augmentation is essential, which can be achieved through techniques such as style transformation, generative models, and data compression. However, there is a clear

a) School of Computer Science and Engineering, Kyungpook National University
‡ Corresponding Author : Sang-hyo Park
E-mail: s.park@knu.ac.kr
Tel: +82-53-950-6373
ORCID: https://orcid.org/0000-0002-7282-7686

※ This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2023-00227431, Development of 3D space digital media standard technology).
· Manuscript October 14, 2024; Revised November 28, 2024; Accepted November 29, 2024.

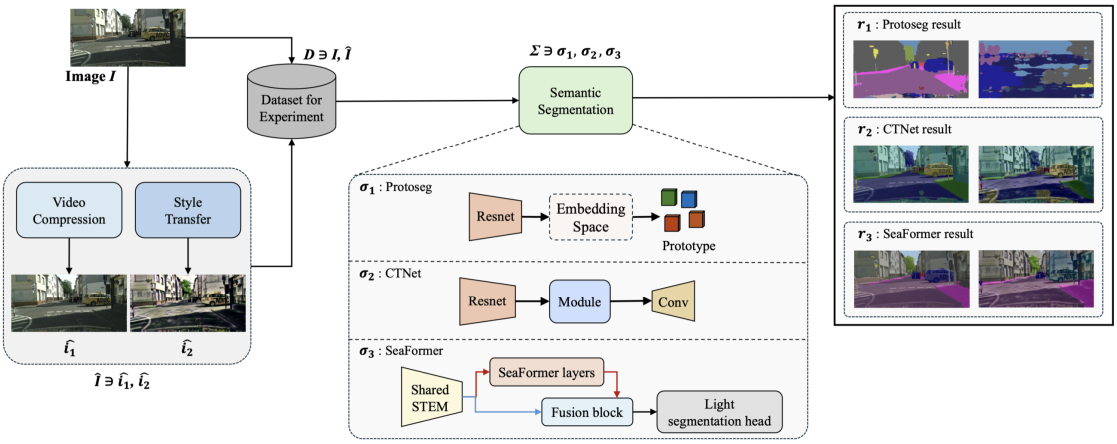


Fig. 1. Proposed framework at the testing stage

domain gap between augmented data and real-world data [3]. Thus, it is essential to examine how the domain gap between these two data may impact the performance of semantic segmentation models.

In this paper, we aim to explore the vulnerabilities of semantic segmentation models when applied to augmented datasets. The framework of this study follows the structure shown in Figure 1. First, we compare and analyze both real and synthetic datasets to select data for the experiments. Then, we build the experimental datasets using style transfer and video compression techniques. Finally, we apply these datasets to semantic segmentation models trained on real data and analyze the results. Through this process, the study aims to identify the vulnerabilities of models trained on real-world data when applied to synthetic datasets.

Ⅱ. Related Work

1. Semantic Segmentation

In autonomous driving, semantic segmentation plays a crucial role in understanding road scenes. For example, semantic segmentation detects and classifies various semantic classes such as pedestrians, vehicles, and trees within an image, allowing for an understanding of the spatial relationships among scene components. This comprehension enables autonomous driving systems to effectively extract road information and make sophisticated decisions[2].

To achieve a comprehensive understanding of semantic segmentation in road scenes, it is essential to have a dataset that includes a variety of road environments. Datasets capturing real-world road scenes, such as CityScapes[4], CamVid[5], KITTI[6], IDD[7] are primarily used for training semantic segmentation models. These road datasets contain scenes captured under diverse conditions, including varying weather, lighting, times of day, and traffic scenarios. Furthermore, the road datasets include critical classes: vehicles, persons, sidewalks, traffic lights, and traffic signs that must be identified while driving.

However, since these real datasets require a significant amount of time for pixel-level annotation tasks, it is challenging to build new training data for all road scenes. To address these limitations, there has been an increasing effort to utilize synthetic data created with computer graphics [8,9,10] or to augment real-world data with generated objects [11].

This paper applies style transfer techniques to both real and synthetic datasets to create a new dataset with generative characteristics. This new dataset is then used with a semantic segmentation model trained on real road datasets. The peculiarities of the generated data are identified by comparing the segmentation results between the existing datasets and the generated datasets.

2. Data Augmentation

Image data augmentation is a technique that generates multiple datasets from a single dataset to address overfitting issues caused by insufficient or imbalanced training data[12]. With advancements in deep learning, methods utilizing image generation models and style transfer models [13,14,15,16] have been developed to augment data.

Style transfer involves applying the style image to a content image to create a new image[17]. Recently, vision-language models have achieved high performance, leading to active research into transforming the style of content images using only text, without requiring reference style images. Notable examples include the StyleCLIP[15] and the CLIPstyler[16] models, both of which utilize CLIP[18], a text-image embedding model developed by OpenAI. The CLIPstyler exhibits superior transfer performance compared to StyleCLIP and has been utilized in this study.

Additionally, methods that use video game engines to render virtual environments similar to the real world are gaining attention for generating synthetic datasets[8,9,10]. By combining basic blocks provided by game engines, new data can be easily generated, significantly enhancing the visual diversity of the dataset. For instance, Virtual KITTI[9] uses the Unity game engine to create a large-scale virtual urban scene dataset for semantic segmentation.

Another method of augmentation is video compression. Video compression is a technique that reduces the redundancy existing in the data to transmit and store the information at a lower bit rate while maintaining its quality [19]. When compressing videos, the quantization parameter, which determines the loss criteria and indicates the strength of the compression, can be used to create transformation datasets with varying qualities. During the video compression process, the internal distribution of data changes, resulting in data augmentation.

Applying data augmentation techniques in semantic segmentation allows for the generation of semantic annotations for augmented data without additional annotation efforts. This approach offers the advantage of acquiring large volumes of data at a low cost. Moreover, these methods maximize the efficiency of data augmentation and contribute to the development of models with improved performance.

Ⅲ. Proposed Method

1. Prompt Selection for Style Transfer Model

To augment the data, this study employs the CLIPstyler [16] model, which facilitates style transfer using only text,

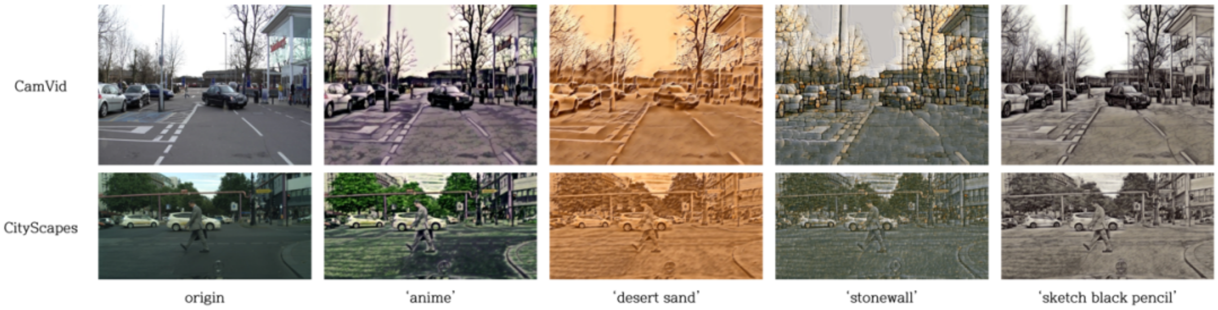


Fig. 2. Comparison of prompt results for CLIPstyler

Table 1. Comparison between the road datasets

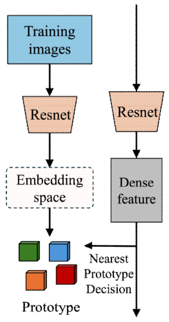
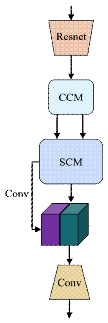
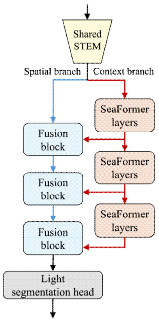
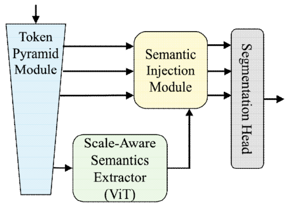
	Real				Synthetic		
Dataset	CamVid	CityScapes	IDD	KITTI	SYNTHIA-SF	Virtual KITTI 2	GTAV
Resolution	960×720	2048×1024	1920×1080	1242×375	1920×1080	1242×375	1914×1052
Number of Data	700	5,000	10,000	200	2,224	21,260	24,966
Classes	32	30	34	-	23	-	19
Used	○	○	×	○	○	×	△
The reason for using or not using	Representative road dataset	Used for model training	Captures unstructured road scenes unlike CityScapes	Primarily used for training autonomous driving	Representative synthetic road dataset	Lack of major classes in road dataset	Appears like real road dataset

to transform existing road datasets. The CLIPstyler provides various prompts such as 'anime', 'stonewall', 'desert sand', and 'sketch black pencil'. The results of style transfer using these prompts are compared and analyzed to select a prompt that preserves the original image information while also imparting generative characteristics. As shown in Figure 2, the 'sketch black pencil' prompt produces grayscale images that lost the color information of the original images. For the 'stonewall' and 'desert sand' prompts, key classes such as vehicles and persons are rendered too blurred, making it difficult to identify the shapes in the original image. Consequently, it has been determined that semantic segmentation models may struggle to perform accurate segmentation. In contrast, the 'anime' prompt preserves the colors and shapes of objects while adding a distinct generative style. Therefore, this study selects the 'anime' prompt for the experiments.

2. Dataset Selection for the Experiment

To select datasets for the experiments, we compared and analyzed four real datasets – CityScapes[4], CamVid[5],

Table 2. A comparison of FCN based models and ViT based models

	FCN based		ViT based	
	ProtoSeg	CTNet	SeaFormer	TopFormer
Year	2022	2021	2023	2022
Backbone	ResNet-101	ResNet-101	SeaFormerB	TopFormer-B
Input size	fixed 1024×512	fixed 1024×512	fixed 1024×512	fixed 1024×512
FLOPs(G)	-	-	3.4	11.2
△Parameters(M)	68.5	-	-	-
Framework				
Used	○	○	○	×

Token Pyramid Module
Semantic Injection Module
Scale-Aware Semantics Extractor (ViT)
Segmentation Head

KITTI[6], and IDD[7]−and three synthetic datasets−SYNTHIA-SF[8], Virtual KITTI 2[9], and GTAV[10]. The results are summarized in Table 1. CityScapes, a benchmark dataset for semantic segmentation, is utilized in the ProtoSeg[20], CTNet[21], and SeaFormer[22] models. KITTI, widely used for training autonomous driving systems, utilized 200 annotated images with semantic segmentation labels for this experiment. CamVid and SYNTHIA-SF are widely recognized road datasets used for training semantic segmentation models and are also included in the experiments. Specifically, CamVid is used for training with real-world data, while SYNTHIA-SF is utilized for training with synthetic data. GTAV, a dataset of urban scenes extracted from a video game, is partially adopted due to its realistic representation of day and night transitions.

On the other hand, IDD contains unstructured road scenes compared to CityScapes, resulting in a significant decrease in segmentation accuracy when models trained on CityScapes are applied to IDD images. Furthermore, Virtual KITTI 2 lacks persons, which are a crucial class for road scenes. Consequently, both IDD and Virtual KITTI 2 are excluded from this study.

3. Models Comparison and Analysis

Semantic segmentation models can be categorized into fully convolutional network (FCN)-based models and vision transformer (ViT)-based models, depending on their backbone architecture. Traditional segmentation models utilize fully convolutional networks for feature extraction; however, with the advent of vision transformers, research has increasingly focused on these models, which have shown superior performance.

In this paper, we compare and analyze the latest FCN-based models, ProtoSeg[20] and CTNet[21], in conjunction with ViT-based models, SeaFormer[22] and TopFormer[23], to select the most suitable models for our experiments. Table 2 provides a summary of the key characteristics of each model.

ProtoSeg, an FCN-based model, effectively captures diverse intra-class variations by utilizing multiple prototypes per class without requiring additional training. CTNet enhances semantic segmentation performance by exploring both spatial and semantic relationships between pixels and channels through its Channel Context Module and Spatial Context Module. On the other hand, ViT-based models like SeaFormer and TopFormer are optimized for mobile devices. SeaFormer captures rich contextual and spatial information through fusion blocks and SeaFormer layers, while TopFormer builds more robust hierarchical features by integrating spatial and semantic information through its Semantics Injection Module.

As shown in Table 2, SeaFormer is approximately three times more efficient than TopFormer in terms of floating point operations per seconds (FLOPs), making it the model used in this study. In addition, ProtoSeg and CTNet were selected to compare the performance of FCN-based semantic segmentation models on the style-transformed dataset generated by the CNN-based model ClipStyler.

CTNet does not follow the class color scheme of CityScapes but instead uses its own color system for segmentation results. Therefore, when comparing the segmentation results of CTNet with those of other models, such as ProtoSeg and SeaFormer, on the same images, the classes in CTNet's results may be represented in colors that differ from the standard CityScapes palette.

Ⅳ. Experiment Result

1. Experimental Setup

The original batch sizes for the ProtoSeg and CTNet models were set at 8 and 16, respectively, while the SeaFormer model utilized 8 GPUs with a batch size of 2.

However, due to resource constraints in this study, the

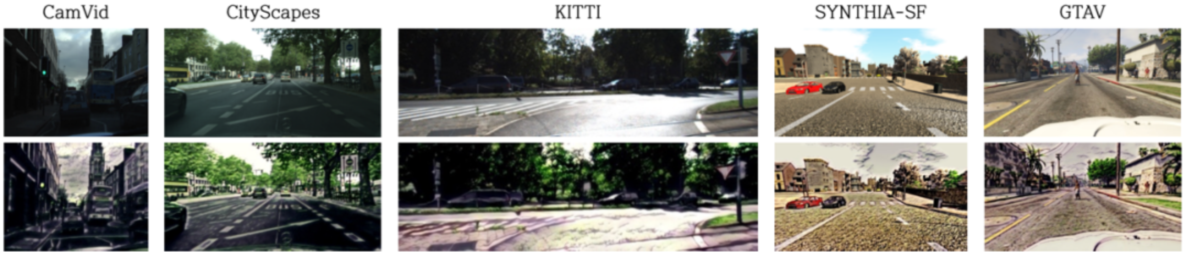


Fig. 3. Road datasets used in the experiment (the upper part shows the video compression, and the lower part shows the style transfer)

batch sizes for ProtoSeg and CTNet are reduced to 4, and the batch size for SeaFormer is set to 8. The parameters for iteration, loss function, and learning rate are all configured to align with those of the original models. The models were implemented using PyTorch. ProtoSeg and SeaFormer are trained on NVIDIA RTX 3080 GPUs, while CTNet is trained on Google Colab's NVIDIA A100-SXM4 GPU.

As shown in Fig. 3, the experimental datasets include style-transferred and video compression datasets. The style transfer is performed using the prompt 'anime.' In addition, to perform video compression using the high efficiency video coding (HEVC) codec, each image from the experimental datasets is converted into a 5 second video. After compression, the videos exhibit characteristics of the YUV420 color space, 4:2:0 chroma subsampling, and an 8 bit depth. The value of QP is set to 10 when observing the performance of ProtoSeg, CTNet, and SeaFormer, and to 30, 40, and 50 when investigating the factors that affect the performance of ProtoSeg. The file sizes before and after compression are presented in Table 3. To apply the semantic segmentation model, the first frame of each video is extracted as an image.

Table 3. File sizes of datasets before and after video compression

Dataset	Resolution	Original file size(KB)	Compressed file size(KB)
CityScapes	1024x512	87.7	46.0
KITTI	1242x375	192	128
SYNTHIA-SF	1920x1080	632	565
GTAV	1914x1052	329	211

2. Results Analysis

In this experiment, the CityScapes, KITTI, CamVid, SYNTHIA-SF, and GTAV datasets, along with their style transformed and video compressed datasets, are applied to ProtoSeg, CTNet, and SeaFormer to compare and analyze semantic segmentation performance, and the results are presented in Table 4. The bold and blue numbers in Table 4 mean the first and second highest performance results for each model, respectively. As seen in the results, the transformed datasets generally presented lower mean intersection over union (mIoU) performance compared to the original datasets. One notable point is the mIoU results of CTNet on the style transformed and video compressed CityScapes: When compared to the results of ProtoSeg and SeaFormer, CTNet shows the most significant mIoU variation between the original dataset and the transformed datasets. This suggests that CTNet is overfitted to the original CityScapes dataset and has difficulty adapting to the distribution changes caused by data transformations. Additionally, the performance of the SeaFormer on the style transfer datasets is significantly superior to that of ProtoSeg and CTNet. While ProtoSeg and CTNet show a considerable performance gap between the original and style transfer datasets, SeaFormer shows a relatively small performance difference of approximately 10% in the mIoU.

Table 4. Quantitative results of the Experiment

		Pixel Accuarcy(%)↑			mIoU(%)↑		
		ProtoSeg	CTNet	SeaFormer	ProtoSeg	CTNet	SeaFormer
Original	CityScapes	**80.9**	**96.0**	82.5	**61.1**	**77.5**	**69.6**
	KITTI	80.6	88.4	84.8	41.6	51.7	56.2
	SYNTHIA-SF	62.1	80.3	**89.9**	15.6	23.9	30.7
	GTAV	66.5	86.7	73.6	34.0	43.7	39.9
Style Transfer	CityScapes	56.3	41.6	70.4	24.7	6.8	53.3
	KITTI	34.2	66.4	71.4	6.3	23.4	47.5
	SYNTHIA-SF	23.7	40.0	73.4	4.7	10.9	20.1
	GTAV	38.2	54.3	62.3	8.6	19.2	25.2
Video Compression	CityScapes	80.1	90.3	82.0	58.2	64.6	66.3
	KITTI	77.6	88.2	83.9	38.6	49.1	54.4
	SYNTHIA-SF	58.6	80.0	89.7	14.4	23.5	30.8
	GTAV	62.3	87.1	73.5	29.6	44.0	39.4

This indicates that SeaFormer is more flexible in adapting to changes in data distribution. The key point to note in the performance results of video compression is the performance of CTNet on the GTAV and SeaFormer on the SYNTHIA-SF. The compressed GTAV and SYNTHIA-SF show slightly higher mIoU compared to the original data for CTNet and SeaFormer, respectively. This result suggests that the video compression process effectively preserved the essential visual features of the data, likely contributing to the performance improvements.

The segmentation results for ProtoSeg, CTNet, and SeaFormer across each dataset can be observed in Figures 4 and 5. While ProtoSeg and CTNet perform relatively well on the original datasets, their performance sig-

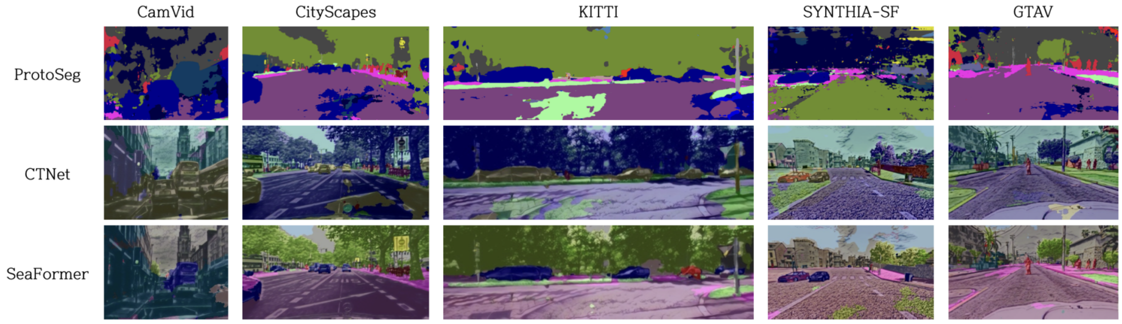


Fig. 4. Qualitative results of the style transfer datasets

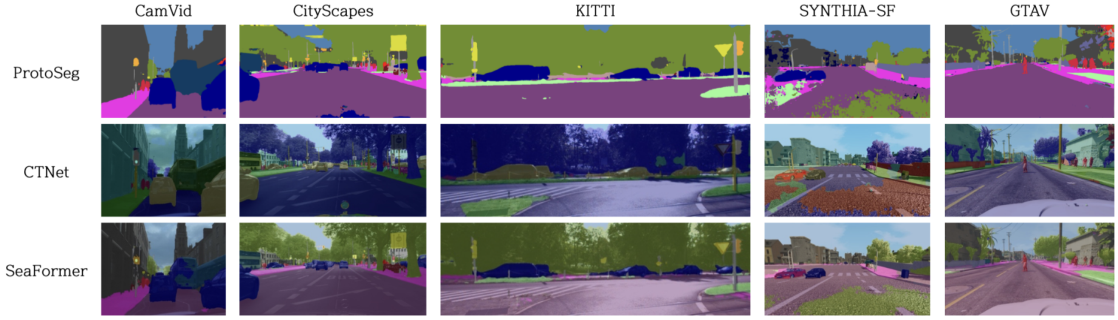


Fig. 5. Qualitative results of the original datasets

nificantly declines on the generated datasets. This decline can be attributed to overfitting to natural datasets, which results in an inability to adapt to the distributional differences present in the style transfer datasets. For instance, the segmentation results of ProtoSeg and CTNet on the CityScapes and KITTI datasets, as shown in Figure 5, reveal that different colors appear in the middle of the road, indicating a failure to accurately identify the 'road' class. Additionally, ProtoSeg does not classify the 'road' as 'vegetation' in the style transformed SYNTHIA-SF dataset, failing to recognize the structure of the original image. This shows the vulnerability of FCN-based models like ProtoSeg and CTNet when applied to generated datasets. Furthermore, although SeaFormer also shows reduced performance on the style transfer datasets compared to the original datasets, it demonstrates clear object segmentation capabilities in specific areas. In particular, in Figure 4, it effectively segments 'buildings' and 'roads' in SYNTHIA-SF, as well as grassy areas on the right sidewalk of GTAV.

We conduct additional research to analyze the factors that cause performance differences and variations in sensitivity in ProtoSeg. We measure the mIoU of the ProtoSeg by setting various quantization parameters for video compression data. As seen in Table 4, the results for video compression with QP set to 10 demonstrate that a low compression rate preserves high data quality, allowing the model to accurately segment object boundaries. However, as shown in Table 5, the mIoU decreases with increasing QP values, recording 55.1 at QP 30, 34.4 at QP 40, and 7.3 at QP 50. This performance degradation occurs because detailed information in the data is gradually lost as the QP increases during the video compression process. These results show that ProtoSeg is sensitive to data quality degradation and struggles to adapt to distribution shifts between the training and testing data.

As mentioned earlier, it is confirmed that ProtoSeg is sensitive and unable to adapt to changes in the internal distribution of data. Therefore, we test whether ProtoSeg would show high adaptability when trained with style-transformed and video-compressed datasets along with CityScapes. The model is trained using CityScapes, an 'anime' style-transformed dataset, and a dataset compressed with a QP value of 40. The performance of the trained model is evaluated on datasets transformed with the 'Desert sand' and 'Sketch black pencil' styles and compressed with QP values set to 30, 40, and 50.

As shown in Table 5, for the style-transformed datasets, the model trained with the transformed datasets achieves mIoU values of 12.1 and 24.8, respectively, improving performance compared to the model trained solely on CityScapes. This shows that the ProtoSeg has been enhanced to withstand changes in color and texture caused by style transformations. Additionally, the model shows stronger performance on video compression dataset transformed using various quantization parameters. Particularly, for the dataset with a QP value of 50, the model trained on the transformed datasets achieves an mIoU of 46.3, which is more than six times higher than the 7.3 achieved by the baseline model. These results confirm its robustness in noise induced environments caused by compression. Therefore, these demonstrate that training with datasets transformed by various styles and quantization parameters enables the model to exhibit stronger performance in handling diverse transformation environments.

Table 5. The mIoU of ProtoSeg

Dataset used for training	Style Transfer		Video Compression		
	"Desert sand"	"Sketch black pencil"	QP 30	QP 40	QP 50
CityScapes	4.8	20.4	55.1	34.4	7.3
CityScapes, style transfer, and video compression	12.1	24.8	71.8	64.2	46.3

Ⅴ. Conclusion

In this paper, we explore the peculiarities of semantic segmentation models when applied to transformed datasets. After constructing new datasets by applying style transformations and video compression to real road datasets, we test models trained on natural datasets. The experimental results show that datasets utilizing data transformation techniques had an impact on the performance of semantic segmentation compared to the original datasets. This study confirms that semantic segmentation models are vulnerable to distortions in color information and object characteristics present in transformed datasets. Additionally, the performance of the compressed dataset is generally similar to that of the original dataset, but shows slightly lower results. We find that ViT-based models are less sensitive to distributional changes and exhibit superior segmentation performance compared to FCN-based models. Through this study, we evaluate how transformed datasets affect model performance and gain insights into how models trained on natural datasets respond to distributional changes. Future research should focus on developing methods to address the vulnerabilities of transformed datasets to further improve semantic segmentation performance.

References

[1] Siam, Mennatullah, et al. “Deep semantic segmentation for automated driving: Taxonomy, roadmap and challenges.” 2017 IEEE 20th international conference on intelligent transportation systems (ITSC), IEEE, Oct 2017.
doi: https://doi.org/10.1109/ITSC.2017.8317714

[2] Meletis, Panagiotis. “Towards holistic scene understanding: Semantic segmentation and beyond.” arXiv preprint arXiv:2201.07734, Jan 2022.
doi: https://doi.org/10.48550/arXiv.2201.07734

[3] Chen, Yuhua, Wen Li, and Luc Van Gool. “Road: Reality oriented adaptation for semantic segmentation of urban scenes.” Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7892-7901, Apr 2018.
doi: https://doi.org/10.48550/arXiv.1711.11556

[4] Cordts, Marius, et al. “The cityscapes dataset for semantic urban scene understanding.” Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3213-3223, June 2016.
doi: https://doi.org/10.1109/cvpr.2016.350

[5] Brostow, Gabriel J., Julien Fauqueur, and Roberto Cipolla. “Semantic object classes in video: A high-definition ground truth database.” Pattern recognition letters 30.2, pp. 88-97, Jan 2009.
doi: https://doi.org/10.1016/j.patrec.2008.04.005

[6] Geiger, Andreas, Philip Lenz, and Raquel Urtasun. “Are we ready for autonomous driving? the kitti vision benchmark suite.” 2012 IEEE conference on computer vision and pattern recognition. IEEE, June 2012.
doi: https://doi.org/10.1109/CVPR.2012.6248074

[7] Varma, Girish, et al. “IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments.” 2019 IEEE winter conference on applications of computer vision (WACV). IEEE, Jan 2019.
doi: https://doi.org/10.1109/WACV.2019.00190

[8] Ros, German, et al. “The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes.” Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3234-3243, June 2016.
doi: https://doi.org/10.1109/cvpr.2016.352

[9] Cabon, Yohann, Naila Murray, and Martin Humenberger. “Virtual kitti 2.” arXiv preprint arXiv:2001.10773, Jan 2020.
doi: https://doi.org/10.48550/arXiv.2001.10773

[10] Richter, Stephan R., et al. “Playing for data: Ground truth from computer games.” Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. Springer International Publishing, 2016.
doi: https://doi.org/10.1007/978-3-319-46475-6_7

[11] Abu Alhaija, Hassan, et al. “Augmented reality meets computer vision: Efficient data generation for urban driving scenes.” International Journal of Computer Vision, Vol 126, pp. 961-972, Mar 2018.
doi: https://doi.org/10.1007/s11263-018-1070-x

[12] Shorten, Connor, and Taghi M. Khoshgoftaar. “A survey on image data augmentation for deep learning.” Journal of big data 6.1, pp. 1-48, July 2019.
doi: https://doi.org/10.1186/s40537-019-0197-0

[13] Ramesh, Aditya, et al. “Hierarchical text-conditional image generation with clip latents.” arXiv preprint arXiv:2204.06125 1.2: 3, Apr 2022.
doi: https://doi.org/10.48550/arXiv.2204.06125

[14] Saharia, Chitwan, et al. “Photorealistic text-to-image diffusion models with deep language understanding.” Advances in neural information processing systems 35, 36479-36494, May 2022.
doi: https://doi.org/10.48550/arXiv.2205.11487

[15] Patashnik, Or, et al. “Styleclip: Text-driven manipulation of stylegan imagery.” Proceedings of the IEEE/CVF international conference on computer vision, pp. 2085-2094, Mar 2021.
doi: https://doi.org/10.48550/arXiv.2103.17249

[16] Kwon, Gihyun, and Jong Chul Ye. “Clipstyler: Image style transfer with a single text condition.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18062-18071, Mar 2022.

doi: https://doi.org/10.48550/arXiv.2112.00374

[17] Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge. “Image style transfer using convolutional neural networks.” Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2414-2423, June 2016.
doi: https://doi.org/10.1109/cvpr.2016.265

[18] Radford, Alec, et al. “Learning transferable visual models from natural language supervision.” International conference on machine learning. PMLR, 139:8748-8763, Feb 2021.
doi: https://doi.org/10.48550/arXiv.2103.00020

[19] Apostolopoulos, John G. “Video compression.,” Springer, MIT 6.344, 2004.

[20] Zhou, Tianfei, et al. “Rethinking semantic segmentation: A prototype view.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2582-2593, Mar 2022.
doi: https://doi.org/10.48550/arXiv.2203.15102

[21] Li, Zechao, et al. “CTNet: Context-based tandem network for semantic segmentation.” IEEE Transactions on Pattern Analysis and Machine Intelligence 44.12, pp. 9904-9917, Dec 2021.
doi: 10.1109/TPAMI.2021.3132068

[22] Wan, Qiang, et al. “SeaFormer++: Squeeze-enhanced axial transformer for mobile visual recognition.” arXiv preprint arXiv:2301.13156, Feb 2023.
doi: https://doi.org/10.48550/arXiv.2301.13156

[23] Zhang, Wenqiang, et al. “Topformer: Token pyramid transformer for mobile semantic segmentation.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12083-12093, Apr 2022.
doi: https://doi.org/10.48550/arXiv.2204.05525

Introduction Authors

Chaewon Moon

- 2021 ~ Present : Undergraduate Student, School of Computer Science and Engineering, Kyungpook National University
- ORCID : https://orcid.org/0009-0001-3492-3626
- Research interests : Computer Vision, Deep Learning

Dong-hwi Kim

- 2021 : B.S. in Information, Communications, Electronics & Engineering from Daejeon University
- 2023 : M.S. in Computer Science and Engineering from Kyungpook National University
- 2023 ~ Present : Ph.D. student in the School of Computer Science and Engineering from Kyungpook National University
- ORCID : https://orcid.org/0000-0002-5188-8834
- Research interests : Computer vision, Deep Learning, Video compression, Generative Model

Dabin Kang

- 2024 : B.S. in Computer Science and Engineering from Kyungpook National University
- 2024 ~ Present : M.S. student in the School of Computer Science and Engineering from Kyungpook National University
- ORCID : https://orcid.org/0009-0000-8242-797X
- Research interests : multi-modal learning, text-to-video retrieval, video question & answering, knowledge distillation, 3D scene understanding

Introduction Authors

Sang-hyo Park

- 2011 : B.S. in Computer Engineering from Hanyang University
- 2017 : Ph.D. in Computer Science from Hanyang University
- 2017 ~ 2018 : Postdoctoral position, Korea Electronics Technology Institute (KETI)
- 2018 : Research Fellow, Yonsei University
- 2019 ~ 2020 : Postdoctoral position, Ewha Womans University
- 2020 ~ Present : Associate Professor with the School of Computer Science and Engineering, Kyungpook National University
- ORCID : https://orcid.org/0000-0002-7282-7686
- Research interests : VVC, Encoding/Decoding Complexity, Omnidirectional Video, Deep Learning, Generative Model