Depth-aware Style Transfer for Robust Data Augmentation in Object Detection

Ian Ryu, Jong-Seok Lee
School of Integrated Technology, Yonsei University
BK21 Graduate Program in Intelligent Semiconductor Technology
Incheon, Republic of Korea
{ianryu, jong-seok.lee}@yonsei.ac.kr

***Abstract*—Data augmentation is a crucial technique for improving the robustness and generalization of deep learning models, particularly in the challenging domain of object detection. However, many traditional approaches such as geometric transformations, color jittering, and noise injection fail to replicate realistic and context-aware environmental variations. This paper presents a depth-aware style transfer approach that leverages depth maps to guide neural style transfer (NST), enabling spatially consistent augmentation. By modulating the style application intensity according to object depth, our method generates augmented data that better preserves semantic structure. Experiments conducted on the Cityscapes dataset using Faster R-CNN demonstrate that the proposed method improves mean average precision (mAP) over conventional style-transfer-based augmentation methods. We also provide an ablation study to highlight the contribution of depth guidance and discuss limitations and potential future directions.**

***Index Terms*—Data Augmentation, Neural Style Transfer, Depth Information, Object Detection, Robustness**

I. Introduction

Object detection is a fundamental task in computer vision that aims to identify and localize objects of interest within an image or video by assigning category labels and bounding boxes. It serves as a cornerstone for numerous applications, including autonomous driving, surveillance, robotics, and medical imaging, where accurate and reliable detection of objects is essential for higher-level decision making. Given its importance, significant research has been devoted to designing detectors such as Faster R-CNN [1] and YOLO [2], which have achieved remarkable progress in both accuracy and efficiency.

Despite these advancements, object detection models must operate robustly under diverse conditions such as weather changes, lighting variations, and scene clutter. Data augmentation is an effective way to address this challenge, which aims to expose the models to training samples of diverse conditions, thereby improving generalization. While conventional augmentation techniques such as flipping, rotation, scaling, and synthetic corruptions have proven beneficial, they are limited in their ability to capture realistic environmental changes that reflect both visual appearance and spatial context.

Neural style transfer [3] enables the synthesis of images by combining the structural content of one image with the visual style of another, which can be used as an advanced approach for data augmentation. However, applying style transformations uniformly across the image may distort the underlying geometry, resulting in unrealistic textures on foreground objects or background elements. In this work, we address this shortcoming by incorporating depth information into the style transfer process, enabling depth-aware modulation of style intensity.

II. Related Work

Advanced augmentation methods such as AutoAugment [4], AugMix [5], and OA-DG [6] improve robustness by algorithmically selecting or mixing transformations. Style-based augmentations have emerged as an effective way to simulate domain shifts, but they often overlook geometric consistency. AdaIN [7] accelerates arbitrary style transfer by aligning feature statistics, and monocular depth estimation methods [8] provide per-pixel geometry cues useful for preserving spatial relationships in augmented images.

In object detection, robustness benchmarking [9] has shown that augmentations simulating real-world variations can significantly reduce performance degradation. Our approach builds upon these insights by directly integrating depth priors into the augmentation process.

III. Proposed Method

A. Concept and Motivation

Our goal is to produce augmented datasets that are not only visually diverse but also spatially coherent. To achieve this, we extract depth maps from the original images and modulate the style intensity based on relative depth. This ensures that style patterns are applied more strongly to distant background regions while preserving texture fidelity on closer, foreground objects that are critical for detection accuracy.

B. Depth-aware AdaIN Transformation

Given content features F_c^l and style features F_s^l at CNN layer l, and depth weights w_d derived from the normalized depth map, we define the transformation as:

$$\hat{F}^l = w_d \cdot \sigma(F_s^l) \left(\frac{F_c^l - \mu(F_c^l)}{\sigma(F_c^l)} \right) + \mu(F_s^l)$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ denote mean and standard deviation. Depth weighting prevents over-stylization of important object regions, thereby preserving their shapes and boundaries.

C. Implementation Details

We implement our pipeline in PyTorch, utilizing pre-computed depth maps from the Cityscapes dataset [10]. The detector is Faster R-CNN [1] with a ResNet-50 backbone, trained with stochastic gradient descent. Style images are drawn from multiple domains to simulate diverse environmental styles.

IV. Experiments

A. Experimental Setup

We evaluate on the Cityscapes validation set, comparing our method against StylizedAug [9], AutoAugment [4], OA-DG [6], and AugMix [5]. We also perform an ablation study to quantify the effect of depth guidance.

B. Quantitative Results

Table I shows that our method achieves a $+1.87$ mAP improvement over StylizedAug, demonstrating the advantage of depth-awareness in style-based augmentation.

TABLE I
COMPARISON WITH STYLE-TRANSFER-BASED AUGMENTATION

Method	mAP
StylizedAug	36.30
Ours	38.17

Table II presents a broader comparison. While OA-DG [6] and AutoAugment [4] achieve higher overall mAP, our method remains competitive and outperforms the other style-transfer-only baselines which includes StylizedAug [9].

TABLE II
MAP COMPARISON WITH STATE-OF-THE-ART METHODS

Method	mAP
AutoAug	42.40
OA-DG	43.40
AugMix	39.50
StylizedAug	36.30
Ours	38.17

C. Qualitative Results

Fig. 1 shows that our augmentation maintains realistic object boundaries and scene geometry while altering visual style.

D. Ablation Study

When depth guidance is disabled, mAP drops by more than 1.5 points, confirming that spatially-aware stylization contributes to performance gains. We also observe that overly aggressive style application can harm detection accuracy, underscoring the importance of depth-modulated blending.

Fig. 1. Example detection results on Cityscapes: (top) original image, (bottom) depth-aware style-transferred image with detection output.

V. Conclusion and Future Work

We introduced a depth-aware style transfer method for robust object detection data augmentation. By combining the strengths of neural style transfer and depth-guided modulation, our approach produces realistic and spatially coherent augmentations. Future work includes integrating our method with other augmentation frameworks (e.g., AugMix) and evaluating on diverse datasets to further validate generalization.

Acknowledgment

This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00453301) and in part by Institute of Information & communications Technology Planning & Evaluation (IITP) under 6G Cloud Research and Education Open Hub (IITP-2025-RS-2024-00428780) grant funded by the Korea government (MSIT).

References

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2015, pp. 91–99.

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 779–788.

[3] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 2414–2423.

[4] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation policies from data," in *CVPR*, 2019, pp. 113–123.

[5] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "Augmix: A simple data processing method to improve robustness and uncertainty," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020.

[6] W. Lee, D. Hong, H. Lim, and H. Myung, "Object-aware domain generalization for object detection," in *AAAI Conference on Artificial Intelligence*, vol. 38, no. 4, 2024, pp. 2947–2955.

[7] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 1501–1510.

[8] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, vol. 44, no. 3, pp. 1623–1637, 2020.

[9] C. Michaelis, B. Mitzkus, R. Geirhos, E. Rusak, O. Bringmann, A. S. Ecker, M. Bethge, and W. Brendel, "Benchmarking robustness in object detection: Autonomous driving when winter is coming," *arXiv preprint arXiv:1907.07484*, 2019.

[10] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 3213–3223.