일반논문 (Regular Paper)
방송공학회논문지 제31권 제1호, 2026년 1월 (JBE Vol.31, No.1, January 2026)
https://doi.org/10.5909/JBE.2026.31.1.162
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

희귀 클래스 인식을 통한 LLM 기반 캡션 확장 기법 – 객체 검출용 확산 데이터셋 응축

땀 린 트란[a)], 배 성 호[a)‡]

LLM-based Rarity-Aware Caption Expansion for Diffusion Dataset Condensation in Object Detection

Linh Tam Tran[a)] and Sung-Ho Bae[a)‡]

요 약

Dataset Condensation(DC)은 대규모 데이터셋을 고신호·저용량의 합성 데이터로 압축하여, 극히 적은 데이터로도 모델 학습을 가능하게 하는 기법이다. 최근 이미지 분류 분야에서는 텍스트-이미지 확산 모델과 결합한 DC 접근법이 성과를 보였으나, 이를 객체 검출(Object Detection)에 그대로 적용할 경우 한계가 드러난다. 캡션 기반의 짧은 프롬프트는 장면 정보를 충분히 기술하지 못하며, 그 결과 클래스 불균형이 증폭되어 확산 모델이 빈도가 높은 클래스에 편향된 이미지를 생성한다. 이는 객체 밀도가 낮고, 공간적 커버리지가 제한적이며, 객체 간 관계 구조가 약한 합성 데이터로 이어진다. 본 논문에서는 이러한 문제를 해결하기 위해 RCE(Rarity-guided Caption Expansion) 프레임워크를 제안한다. RCE는 먼저 학습 데이터셋으로부터 클래스 희소도를 추정하고, 경량의 객체 공출현(co-occurrence) 사전 정보를 구성한다. 이후 대규모 언어 모델(LLM)을 활용하여 장면의 일관성을 유지하면서 희소 클래스 객체를 프롬프트에 선택적으로 확장 삽입함으로써, 의미적 엔트로피를 증가시키고 클래스 편향을 완화한다. COCO 데이터셋에서 극단적인 압축 예산 조건 하에 실험한 결과, RCE는 희소 클래스 평균 정밀도(AP)를 최대 +18.8(4.6 → 23.4) 향상시켰으며, scissors(+1.9), microwave(+6.4), fire hydrant(+16.0), stop sign(+11.7) 등 데이터가 부족한 카테고리 전반에서 일관된 성능 개선을 달성하였다.

Abstract

Dataset Condensation compresses big datasets into concise, high-signal synthetic subsets so models learn with far less data. While recent work couples DC with text-to-image diffusion for image classification, directly porting this recipe to object detection exposes a bottleneck: short, caption-derived prompts under-specify scenes and amplify class imbalance, so diffusion preferentially samples head classes and produces images with low object density, limited spatial coverage, and weak relational structure. We introduce RCE, a Rarity-guided Caption Expansion framework that counteracts this bias. RCE first estimates class rarity from dataset annotations and derives a lightweight co-occurrence prior; it then uses a Large Language Model (LLM) to insert scene-compatible rare objects into prompts, increasing semantic entropy while preserving consistency. On COCO under extreme compression budgets, RCE improves rare-class AP by up to +18.8 (4.6 → 23.4) and consistently boosts scarce categories such as scissors (+1.9), microwave (+6.4), fire hydrant (+16.0), and stop sign (+11.7).

Keyword : Dataset Condensation, Object Detection, Large Language Model

Ⅰ. Introduction

Deep neural networks (DNNs) are now the standard approach for many computer vision problems [1,2,3,4]. Yet training them well typically requires large datasets and substantial GPU time. Dataset Condensation (DC) [5,6,7] addresses this by compressing a massive dataset into a much smaller synthetic set, cutting both computation and storage. The goal is for the condensed data to deliver performance close to the original, enabling efficient experimentation under limited resources.

Because classification DC methods depend on bi-level gradient matching that jointly tunes synthetic data and the model [5,6,7], they are costly to run. When transferred to object detection—where images are high-resolution and instance counts drive memory and time—the expense grows sharply. This practical barrier helps explain why detection has seen far less work than classification.

Progress on DC for object detection remains limited. Early work, such as DCOD [8], follows a two-stage pipeline: pretrain a detector on the full dataset, then synthesize images via model inversion. In contrast, recent approaches, namely UniDD [9], leverage text-to-image (T2I) diffusion to generate training data. However, the prompts are typically short and mirror the original captions, which under-specify the scene and amplify class imbalance: common objects are oversampled while rare categories are undersampled. As a result, tail categories are systematically underrepresented in the synthesized set.

To address this challenge, we propose a simple but effective rarity-guided caption expansion that first estimates class rarity from dataset annotations and then uses a Large Language Model (LLM) to insert scene-compatible rare objects, attributes, and relations into prompts, thereby increasing tail coverage and restoring layout diversity while preserving scene consistency. On MS COCO [10], our method improves overall mAP and yields larger gains on rare classes, demonstrate the effectiveness of the proposed method.

Ⅱ. Related works

Dataset Condensation: First proposed by Wang et al. [5], DC aims to generate a small synthetic dataset on which a model can train to near full-data performance. Over time, researchers have explored diverse objectives—trajectory matching [11], distribution matching [12], representative matching [13], and gradient matching [6].

Diffusion-based methods: While effective for image classification, extending image-domain DC methods to object detection is computationally prohibitive: detectors require high-resolution inputs and multi-object layouts, which sharply increase memory and optimization cost. To avoid heavy image-space optimization, recent work turns to text-to-image (T2I) diffusion to synthesize training data; for example, UniDD [9] uses Stable Diffusion [14] to generate condensed images. However, prompting T2I with short, caption-derived text under-specifies the scene and amplifies head-class bias (e.g., frequent “person”), leaving rare categories underrepresented.

To mitigate this, we propose a rarity-guided caption expansion: estimate class rarity from dataset annotations, then use an LLM to inject scene-compatible rare objects (with attributes/relations) into the prompt, and finally synthesize images via T2I. Despite its simplicity, this strategy in-

a) 경희대학교 컴퓨터공학과(Department of Computer Science and Engineering, Kyung Hee University)
‡ Corresponding Author : 배성호(Sung-Ho Bae)
E-mail: shbae@khu.ac.kr
Tel: +82-31-201-2593
ORCID: https://orcid.org/0000-0003-2677-3186
※ This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support programs (IITP-2025-RS-2023-00258649), (IITP-2025-RS-2023-00259004), supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation)
· Manuscript December 18, 2025; Revised January 5, 2026; Accepted January 5, 2026.

creases tail coverage and strengthens rare-class learning.

Ⅲ. Proposed Method

1. Empirical Evidence on Rare Classes

To highlight the long-tail imbalance, we compute per-class instance frequencies on COCO [10] by counting every annotated bounding box in 'instances_train2017.json', and plot their distribution in Figure 1.

Figure 1 shows that the person class appears far more frequently than others, whereas tail classes such as hair drier are severely underrepresented. When one class dominates others, head classes drive optimization while tails are under optimization. The result is collapsed tail recall, especially under tiny condensed-data budgets.

2. Rarity-guided Caption Expansion (RCE)

To remedy this issue, we explicitly inject rare object classes into the caption while preserving semantic consistency with the original scene via an LLM. RCE consists of three steps:

1) Rare-class computation: We compute class frequencies over the training set by counting the total bounding boxes for the given class. Based on these frequencies, we select the bottom-k classes, which we treat as rare classes.
2) Rarity-conditioned prompting: We design a structured LLM prompt p that instructs the model to incorporate the selected rare classes o into the caption while maintaining scene coherence.
   Example prompt used in our method:
   "Original Caption: <base-caption>
   Here is a list of possible objects to consider:
   <rare-classes>
   Select the most contextually appropriate objects and incorporate as many as fit naturally."
3) Caption-rewrite: Given a caption c (randomly selected from the dataset) and the selected rare classes o, we prompt an LLM to expand the caption as:

$$c^* = M(c, o, p) \tag{1}$$

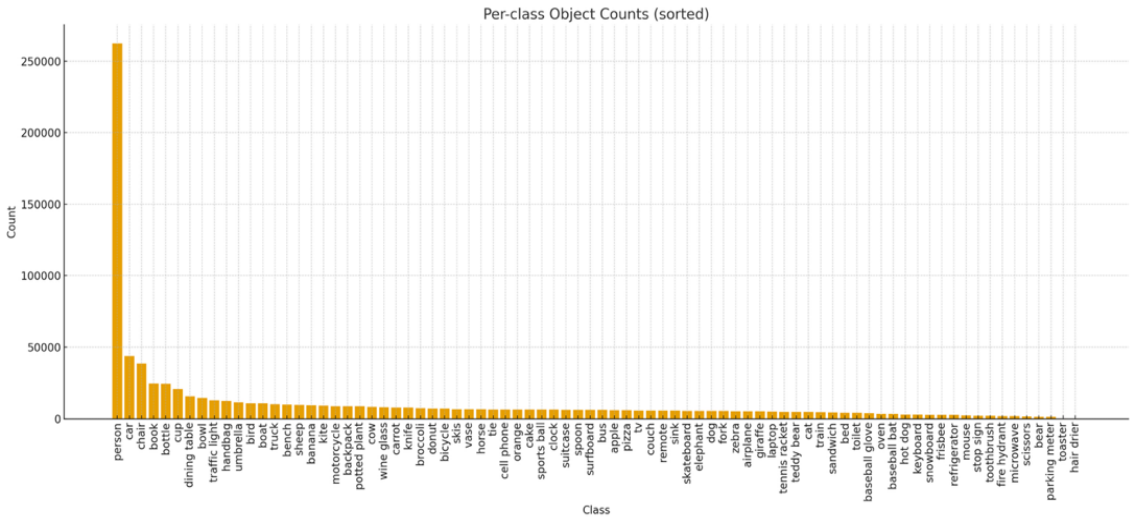

그림 1. MS COCO 데이터셋의 클래스별 개수
Fig. 1. Class count on MS COCO

By adding rare classes to the caption, we increase the sample of rare classes, which improves the learning for these classes.

3. Synthetic Data Generation Pipeline

Our pipeline proceeds as follows. We first sample a seed caption from the training captions. Using dataset statistics, we select a bottom-*k* set of rare classes and prompt an LLM to expand the seed into an enriched caption by inserting those classes with brief attributes/relations. We then condition Stable Diffusion on the enriched caption to synthesize an image. Finally, a strong pre-trained detector is run on the synthetic image to produce pseudo-bounding boxes, yielding a labeled synthetic pair for training.

Ⅳ. Experiments

1. Experiment setting

Dataset: We perform synthesis using MS COCO [10], which has 118,287 images for training and 5,000 images for validation. There are 80 classes in this dataset. COCO comprises everyday scenes with multiple objects per image, diverse viewpoints, and a pronounced long-tail frequency distribution across classes (e.g., frequent “person” vs. rare categories).

Implementation detail: For caption expansion, we adopt GPT-4.1 as the language model. We use Stable Diffusion v3 [14] as the image generation model. For pseudo labeling, we employ a pretrained Faster R-CNN [15].

Architecture: We adopt Faster R-CNN [15] as the detector for training and evaluation.

Storage budget and evaluation metrics: To stress-test learning from condensed data, we adopt extreme training budgets: 0.25%, 0.5%, and 1% on COCO and report the performance with mean Average Precision (mAP).

Comparison method: We compare our method to coreset selection methods such as Random selection [16], K-Center [17], Herding [18], and synthesizing based such as UniDD [9].

2. Main results

2.1 Overall performance

Table 1 reports COCO mAP under extreme budgets (0.25/0.5/1%). Coreset selection (Random, K-Center, Herding) performs poorly compared to diffusion-based synthesis (UniDD, Ours): at 0.25%, coreset methods reach 0.4 – 0.5 mAP, whereas UniDD attains 4.5±0.3. Our rarity-guided expansion further improves low-budget performance to 5.3±0.5 (+0.8 over UniDD; +17.8% relative). At 0.5%, we obtain 8.0±0.3 (+0.9, +12.7% over UniDD). At 1%, our method is 10.5±0.4, slightly below UniDD (10.8±0.4), a – 0.3 difference within one standard deviation, indicating parity at this budget.

표 1. 제안한 방법과 기존 방법 간 성능(mAP) 비교
Table 1. Performance comparison (mAP) between others our method

Ratio	**0.25%**	**0.5%**	**1%**
Random	0.5±0.1	3.7±0.2	7.2±0.8
K-Center	0.4±0.2	3.2±0.5	6.1±0.3
Herding	0.5±0.1	3.5±0.3	6.7±0.4
UniDD	4.5±0.3	7.1±0.4	10.8±0.4
Ours	**5.3±0.5**	**8.0±0.3**	**10.5±0.4**

2.2 Result on rare class

To demonstrate the effectiveness of our rare class boost-

표 2. RCE 적용 시 성능 비교 (비율 = 0.25%)
Table 2. Performance comparison when applying RCE. Ratio = 0.25%

Class name	Parking Meter	Bear	Scissors	Microwave	Fire hydrant	Stop sign
Baseline	0.0	4.6	0.8	4.6	11.8	21.3
+ RCE	0.1	23.4	2.7	11.0	27.8	33.0

ing, we report the mAP for these classes. We select several rare classes and report them in Table 2. As shown in Table 2, enriching the base captions with rarity-guided additions consistently improves detection on tail categories. Relative to using unedited captions (baseline), our method raises AP for rare classes. The results indicate that injecting scene-compatible rare objects and relations yields more informative positives and richer co-occurrence patterns, thereby improving tail recall and localization quality.

3. Ablation study

3.1 Experiment using different k value

We vary k, the number of rare categories inserted per caption. As Figure 2 shows, increasing k generally raises overall mAP by broadening tail coverage and co-occurrence diversity. However, per-class AP is non-monotonic: while some classes continue to benefit, others experience performance degradation when k becomes large. as the candidate pool expands, the LLM increasingly prioritizes the most contextually compatible objects with the input caption. Consequently, certain rare classes may be selected less frequently or omitted altogether, leading to reduced AP for those classes. This observation suggests that overly large k can weaken class-specific effectiveness, highlighting a trade-off between increasing candidate diversity and maintaining balanced rare-class coverage. Nevertheless, our proposed component effectively improves mAP for tail classes.

3.2 Example of expanded caption

Table 3 presents qualitative examples of captions expanded by RCE. Given a base caption and a list of rare classes (hair drier, toaster, parking meter, bear, scissors), the LLM does not simply append all the classes. Instead,

표 3. 기본 캡션과 확장된 캡션의 비교
Table 3. Comparison of base caption and its expended one

Rare classes	Base caption	Expanded caption
hair drier, toaster, parking meter, bear, scissors	A knife that is inside of an apple.	A knife is lodged inside an apple on a kitchen table beside a pair of scissors, a toaster, and a hair drier.
	A train engine with carts pulling into a station.	A train engine with carts pulls into a station as a parking meter stands nearby and a bear waits curiously on the platform.

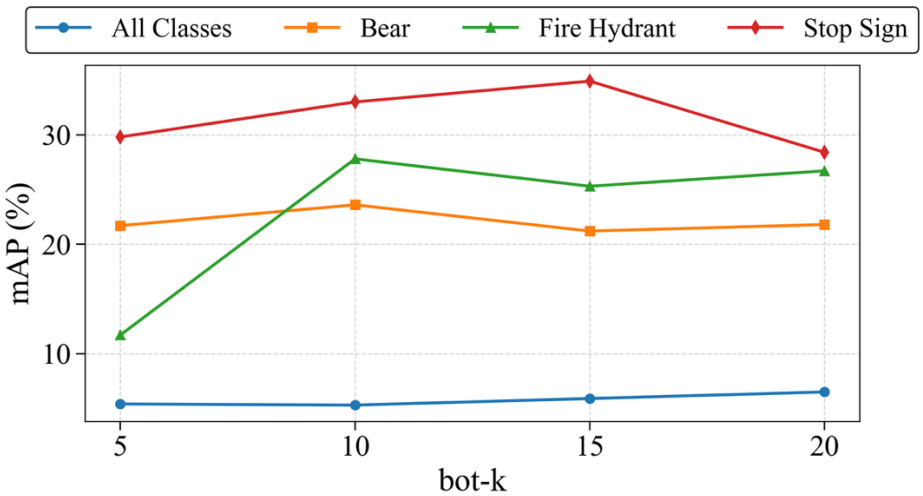

그림 2. 희귀 클래스 수가 성능에 미치는 영향
Fig. 2. Effect of the number of rare classes on performance

it selects those that are semantically compatible with the scene and integrates them into a coherent description. For instance, in the first row, scissors, toaster, and hair drier are added to a kitchen scene, while bear and parking meter are omitted as they are unlikely to appear indoors. Conversely, in the second row, bear and parking meter are used in an outdoor station scene, whereas kitchen-related objects are excluded. This demonstrates that RCE enriches captions with rare classes while preserving scene plausibility.

Ⅴ. Conclusion

We presented a simple yet effective approach to long-tail bias in dataset condensation for object detection. Our rarity-guided caption expansion estimates class rarity from training annotations and uses an LLM to inject scene-compatible rare objects, attributes, and relations into prompts before T2I synthesis. This increases tail coverage and co-occurrence/layout diversity in the condensed set, yielding consistent gains in tail AP—and competitive overall mAP—under extreme compression budgets on COCO. Our method, rare-class enhancement via LLMs, is especially useful for datasets with very limited samples for classes that are hard to capture or rarely occur in real-world scenarios such as emergency equipment, uncommon wildlife, industrial failures, or infrequent safety-critical events.

참 고 문 헌 (References)

[1] K. He, X. Zhang, S. Ren and J. Sun, “Deep Residual Learning for Image Recognition,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778.
doi: https://doi.org/10.1109/CVPR.2016.90

[2] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 779-788.
doi: https://doi.org/10.1109/CVPR.2016.91

[3] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2015), arXiv preprint arXiv:1409.155.
doi: https://doi.org/10.48550/arXiv.1409.155

[4] Yamanaka, J., Kuwashima, S., Kurita, T.: “Fast and accurate image super resolution by deep cnn with skip connection and network in network”, Neural Information Processing, pp.217-225.
doi: https://doi.org/10.1007/978-3-319-70096-0_23

[5] Wang, T., Zhu, J., Torralba, A., Efros, A.A.: Dataset distillation. CoRR abs/1811.10959 (2018).
doi: https://doi.org/10.48550/arXiv.1811.10959

[6] Zhao, B., Mopuri, K.R., Bilen, H.: Dataset condensation with gradient matching. In: International Conference on Learning Representations (2021).
doi: https://doi.org/10.48550/arXiv.2006.05929

[7] Zhao, B., Bilen, H.: Dataset condensation with differentiable siamese augmentation. In: Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 12674 – 12685. PMLR (18 – 24 Jul 2021).
doi: https://doi.org/10.48550/arXiv.2102.08259

[8] Ding Qi, Jian Li, Jinlong Peng, Bo Zhao, Shuguang Dou, Jialin Li, Jiangning Zhang, Yabiao Wang, Chengjie Wang, and Cairong Zhao: Fetch and forge: Efficient dataset condensation for object detection. In The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024).
doi: https://doi.org/10.52202/079017-3790

[9] D. Qi et al., “Towards Universal Dataset Distillation via Task-Driven Diffusion,” 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2025, pp. 10557-10566.
doi: https://doi.org/10.1109/CVPR52734.2025.00987

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), Computer Vision – ECCV 2014, pp.740 – 755, Cham, 2014.
doi: https://doi.org/10.1007/978-3-319-10602-1_48

[11] G. Cazenavette, T. Wang, A. Torralba, A. A. Efros and J. -Y. Zhu, “Dataset Distillation by Matching Training Trajectories,” 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 10708-10717.
doi: https://doi.org/10.1109/CVPR52688.2022.01045

[12] B. Zhao and H. Bilen, “Dataset Condensation with Distribution Matching,” 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 2023, pp. 6503-6512.
doi: https://doi.org/10.1109/WACV56688.2023.00645

[13] Y. Liu, J. Gu, K. Wang, Z. Zhu, W. Jiang and Y. You, “DREAM: Efficient Dataset Distillation by Representative Matching,” 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2023, pp. 17268-17278.
doi: https://doi.org/10.1109/ICCV51070.2023.01588

[14] Stability AI. Stable diffusion 3: Next-generation text-to-image generation, 2024.
[15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), Advances in Neural Information Processing Systems, volume 28. Curran Associates, Inc., 2015.
doi: https://doi.org/10.48550/arXiv.1506.01497
[16] Yutian Chen, Max Welling, and Alex Smola. Super-samples from kernel herding. In UAI, UAI'10, pp. 109–116, Arlington, Virginia, USA, 2010. AUAI Press. ISBN 9780974903965.
doi: https://doi.org/10.48550/arXiv.1203.3472
[17] Ivor W. Tsang, James T. Kwok, and Pak-Ming Cheung. Core vector machines: Fast svm training on very large data sets. Journal of Machine Learning Research, 6(13):363–392, 2005. URL http://jmlr.org/papers/v6/tsang05a.html.
[18] Francisco M. Castro, Manuel J. Mar´ın-Jimenez, Nicol ´ as Guil, Cordelia Schmid, and Karteek Ala- ´hari. End-to-end incremental learning. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), ECCV, pp. 241–257, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01258-8
doi: https://doi.org/10.48550/arXiv.1807.09536

저 자 소 개

Linh Tam Tran

- Sep. 2009 ~ Feb. 2014 : He received the bachelor's degree from the Department of Computer Science and Engineering, Ho Chi Minh City University of Technology, Vietnam
- Mar. 2016 ~ Feb. 2018 : He received the the M.S. degree from Hongik University, Seoul, South Korea
- Sep. 2020 ~ Present : He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, Kyung Hee University, Yongin, South Korea.
- ORCID : https://orcid.org/0000-0002-9699-1747
- 주관심분야 : Dataset Condensation

배 성 호

- Mar. 2004 ~ Feb. 2011 : Bachelor's degree in Department of Computer Engineering and Electronic Engineering (dual majors) at Kyung Hee University, South Korea
- Feb. 2011 ~ Aug. 2016 : Ph.D. in Department of Electrical Engineering at KAIST, South Korea
- Jul. 2016 ~ Aug. 2017 : Postdoc. Associate in MIT Computer Science and Artificial Intelligence Laboratory (CSAIL)
- Sept. 2017 ~ Present : Associate Professor at Kyung Hee University in school of Computing, South Korea
- ORCID : https://orcid.org/0000-0003-2677-3186
- 주관심분야 : Inverse problems in image processing, video compression, model compression, generative AI