CLP: A Real-World Dataset of Contaminated Lens Protectors for Robust Semantic Segmentation

Sungyong Park[1,2*], Sooyoung Choi[1,2*], Hyunseo Koh[1,2], Youngjae Choi[1], Heewon Kim[1,2]
[1]Soongsil University, [2]Kairoba Inc.
https://clp-dataset.github.io

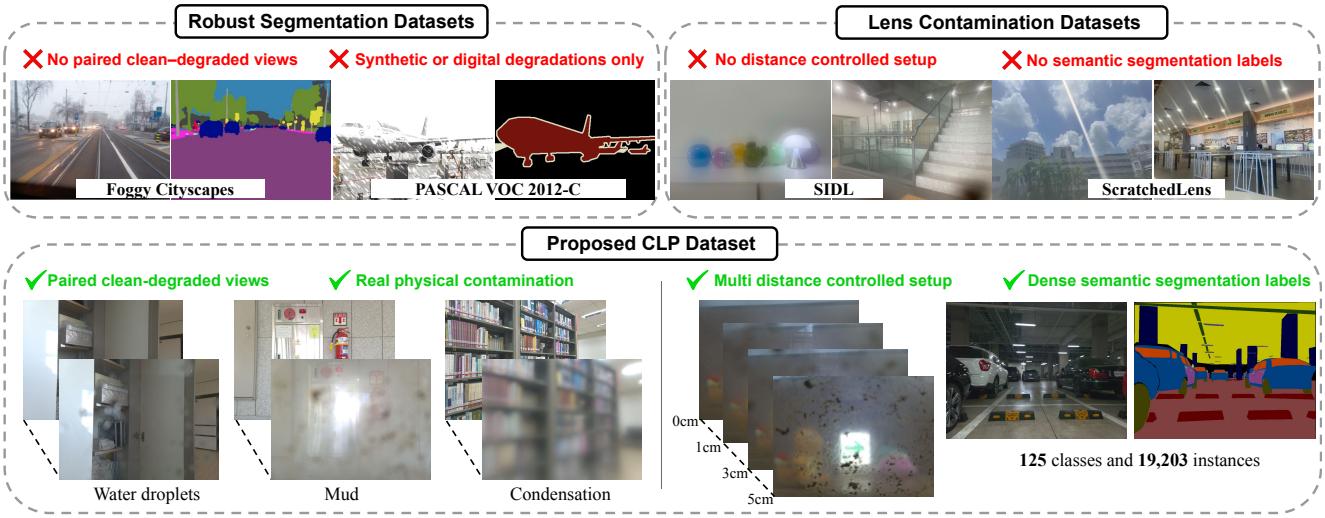

Figure 1. Existing robustness and lens contamination datasets have a limited scope, lacking realistic contamination, aligned pairs, or annotations. CLP offers real physical contamination with paired clean/degraded image pairs, distance control, and dense semantic labels.

Abstract

The reliability of autonomous systems in real-world environments is mainly dependent on the robustness of their visual perception. Although recent studies have advanced the handling of visual degradations, physical contaminants that adhere to the camera lens—such as mud, water droplets, and condensation—remain largely underexplored. To this end, we introduce the CLP (Contaminated Lens Protector) dataset, a real-world benchmark designed to evaluate perception performance under realistic lens-protector contamination. The CLP dataset offers degraded images across multiple types of contamination and various lens-to-protector distances, along with dense semantic segmentation masks and aligned restoration targets. CLP enables robust segmentation and restoration studies in conditions that closely match those encountered by real-world autonomous systems. Experiments analyze strategies to improve perception under contamination with limited data, highlighting the importance of domain generalization, foundation models, data scale, and joint restoration-segmentation pipelines.

* Equal contribution.

1. Introduction

Semantic segmentation has achieved remarkable progress with the advent of large-scale datasets [40, 83] and foundational models [33]. This progress has enabled high-precision scene understanding by capturing rich spatial and semantic cues. For robust perception, recent benchmarks cover digital corruptions [27, 28] and environmental conditions like adverse weather [5, 25, 38, 57, 58]. However, the semantic segmentation of images with a camera lens contaminated by mud or water droplets remains largely unexplored.

The challenge of lens contamination arises from its severe and spatially irregular artifacts (*e.g.*, blur, color shift, occlusion, scattering) and from the difficulty of collecting paired clean–contaminated images. Previous studies have partially explored this issue using real paired data captured with custom hardware [13, 67]. However, these approaches overlook a crucial aspect of real-world lens setups—the physical gap between the camera lens and the lens protector, which exhibits a significantly different degradation pattern.

Digital cameras, smartphones, automotive systems, and humanoid robots typically place a flat and transparent lens

Figure 2. Example scenes from the CLP dataset illustrating lens protector contamination. Each column shows a scene under a different condition: Clean, Mud, Water Droplets, and Condensation. From top to bottom: (1) original (protector-free) scene, (2) degraded version with lens contamination, and (3) corresponding semantic segmentation mask. CLP provides aligned clean/degraded image pairs and dense annotations to facilitate robust segmentation under real-world contamination.

protector (or housing) in front of the camera lens. The gap between the lens and its protector varies across devices to accommodate different needs, such as temperature regulation, shock absorption, and ease of maintenance. However, this gap induces different refraction effects, causing identical contaminants on the protector surface to produce markedly different image degradation patterns, which in turn makes robust perception challenging (See Figure 1).

To understand the effect of this gap on perception, we introduce the **CLP** (**C**ontaminated **L**ens **P**rotectors) dataset, a real-world dataset designed for evaluating perception robustness under lens protector contamination. The CLP dataset provides (1) paired clean and degraded images under mud, water droplets, and condensation on lens protectors; (2) dense semantic segmentation annotations for recognition performance evaluation; and (3) multi-distance captures reflecting realistic lens-to-protector geometries, where visual degradation varies with spacing.

We built a 3D-printed protector holder for mounting replaceable transparent glass plates in front of the lens. This design enables precise geometric control, ensuring consistent scene alignment. Each scene was recorded with a clean protector and with one of three contamination types—mud, water droplets, or condensation—applied in unique patterns at four lens-to-protector distances (0cm, 1cm, 3cm, 5cm), resulting in **4,800** ($600\times2\times4$) degraded images paired with original (protector-free) and clean (contamination-free) references. The CLP dataset further provides dense semantic labels spanning **125** object categories with **19,203** instances, enabling rigorous analysis of recognition performance under realistic sensor degradation.

We conduct comprehensive experiments across diverse learning paradigms and architectures to investigate the impact of methodological choices on robustness across segmentation and restoration tasks. Furthermore, we explore the effects of data scaling and the synergy between restoration and segmentation. This analysis provides an in-depth understanding of the challenges of lens contamination, offering guidance for future research on robust vision systems.

In summary, our contributions are:

- A real-world dataset, CLP, that captures realistic lens protector contamination across multiple types and distances, providing paired clean–contaminated images and dense semantic annotations for 125 categories.
- A data collection framework with a custom 3D-printed holder enabling reproducible captures and precise control of lens–to-protector distance.
- Comprehensive experiments and analysis on diverse segmentation and restoration models, establishing baselines and providing insights for robust perception systems.

Figure 3. Data collection setup and devices equipped with lens protectors. (a) A custom 3D-printed smartphone holder designed to fix glass plates simulating contamination at fixed distances (0cm, 1cm, 3cm, 5cm). (b) Lens protectors are mounted at varying distances in different camera systems to protect the optics from contaminants. (c) By placing contaminated glass between the scene and the camera, our setup captures degraded images with realistic optical distortion and visual degradation.

2. Related Work

Generic Semantic Segmentation. Early semantic segmentation works focused on context modeling with fully convolutional networks [8, 9, 45, 73, 79]. The success of self-attention [64] in modeling long-range dependencies [16, 26, 66, 80] shifted the paradigm toward transformer-based architectures [15, 60, 69, 82]. Mask2Former [11] unified diverse segmentation tasks by reformulating pixel-wise prediction into query-based mask classification. Recently, efficient state-space models [44] and vision foundation models [33, 48, 54] pretrained on massive datasets [14, 20, 40, 47, 72, 83] have achieved strong performance, yet often lack robustness to real-world physical degradations like lens contamination.

Robust Semantic Segmentation. Recent work on robust segmentation follows two principal strategies. Domain Adaptation (DA) [22–24, 49, 62, 65] aligns feature distributions between clean and unlabeled degraded domains, while Domain Generalization (DG) [4, 12, 30, 35, 36, 68, 75] learns domain-invariant representations from multiple sources. Another direction leverages foundation models [10, 37] trained on diverse visual data to improve robustness under common degradations. These paradigms suit lens contamination, which presents unique domain shifts via optical distortion and partial occlusion. In this work, we use CLP to systematically evaluate these methods under lens contamination and outline their strengths and limitations.

Robust Semantic Segmentation Datasets. Real-world robustness datasets [5, 57, 58, 74] capture challenging conditions (*e.g.,* fog, rain, illumination); however, they lack pixel-level alignment between clear and degraded views. To address this, prior work synthesizes artifacts on clean images to obtain scalable aligned pairs, covering weather effects [25, 38, 56], digital corruptions [27, 53], and dirty lens artifacts [63]. However, synthetic artifacts cannot fully capture the complexity of real-world contamination. Another line of work explores controlled physical proxy setups [51]. To obtain aligned pairs with semantic labels, subsequent work [52] re-recorded datasets displayed on a monitor through a water-sprayed glass plate. CLP aims to fill this gap by offering real-world lens contamination with paired clean–degraded images and dense semantic labels, providing a valuable resource for robustness evaluation.

Perceptual Image Restoration focuses on improving the visual quality of degraded images. Deep learning methods based on CNNs [7, 77] and Transformers [34, 76] have established strong baselines. Recently, diffusion-based models [41, 71], state-space models [19], and unified all-in-one architectures [78, 81] have further advanced restoration quality. Several datasets benchmark specific degradations such as noise [1, 50], blur [46, 55], and adverse weather [2, 3]. However, real-world degradations like lens contamination remain largely unexplored. Early efforts [17, 18] addressed lens contamination through physics-based modeling but did not provide aligned image pairs for data-driven restoration. Recent works [13, 39, 67] target physical contaminants yet remain limited. SIDL [13] covers only smartphones, whose fixed configuration prevents systematic variation of factors such as lens-to-protector distance. In contrast, CLP captures realistic lens-protector contamination across multiple distances and diverse environments.

Task-Aware Image Restoration aims to improve degraded inputs for downstream tasks. Early works explored integrating enhancement modules into downstream architectures [29, 42, 43, 59] or employing task-driven losses to guide restoration networks [21, 31, 61]. More recently, UniRestore [6] and EDTR [32] leverage diffusion priors to improve both visual quality and downstream task performance. However, these methods primarily rely on synthetic or unpaired artifacts due to the lack of real-world aligned pairs and pixel-level annotations. CLP addresses this with aligned clean–degraded pairs and dense semantic labels, enabling joint evaluation of restoration and segmentation robustness under real contamination.

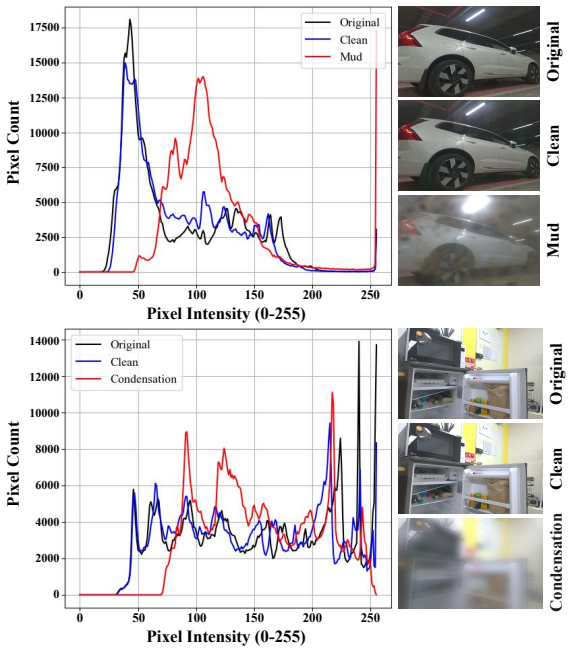

Figure 4. Histogram analysis of contaminated lens protectors. Each plot compares pixel-intensity distributions across the original, clean, and degraded images (Mud, Condensation).

3. CLP Dataset

3.1. Data Collection Setup

Our goal is to collect a diverse set of images simulating lens protector contamination scenarios commonly encountered by cameras deployed in real-world environments. However, collecting such data directly with physical robots (or vehicles) is costly and difficult to scale. Instead, we used smartphone[1] cameras as a proxy and designed a 3D-printed smartphone mount (Figure 3(a)) that holds a thin glass plate at fixed distances (0cm, 1cm, 3cm, 5cm) in front of the camera lens. We vary the distances to reflect the range of lens-to-protector spacings observed in different camera systems and use cases.

3.2. Data Collection Process

To ensure consistent comparisons under varying contamination conditions, we selected scenes (or objects) that remain unchanged throughout data collection. Under these controlled settings, we captured data across diverse environments (*e.g.,* indoor and outdoor, day and night, and various object categories) to reflect the range of real-world conditions. Each scene was captured from one of three camera angles (upward, horizontal, or downward), chosen to correspond to the viewpoints of different camera mounting scenarios. After adjusting the camera to the selected viewpoint, we captured images by moving the contaminated glass plate to four preset distances using our custom holder.

[1]Samsung Galaxy S20 Ultra and iPhone 12 Pro

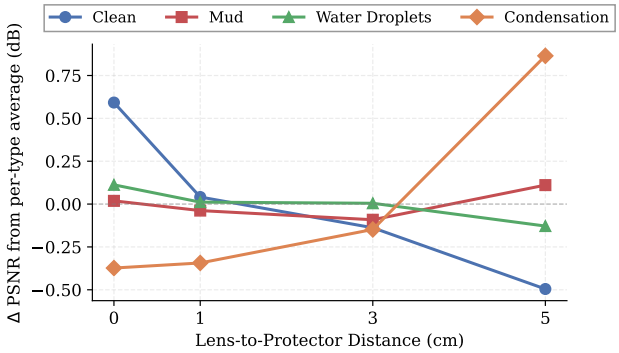

Figure 5. Analysis of lens-to-protector distance effects by contamination type. Each point shows the PSNR difference to the mean PSNR over distances at the same contamination type. Clean and condensation exhibit opposite trends, while mud and water droplets remain stable, indicating that the distance effect varies significantly by contamination type. See the Supplement for visual examples.

3.3. Preprocessing

All images were captured in raw format (4000×3000 pixels) to preserve full sensor fidelity and avoid compression artifacts. The raw data were converted to RGB using an open-source ISP pipeline, calibrated using each image's metadata. After conversion, images were resized to 1000×750 pixels to balance computational efficiency and detail preservation for downstream tasks.

3.4. Contamination Types and Characteristics

The CLP dataset includes images captured under five distinct conditions: one *original* setting and four types of contamination—*clean*, *water*, *condensation*, and *mud*.

Original. Original images are captured without a glass plate in front of the lens. They serve as protector-free references for image restoration and segmentation annotation.

Clean. Clean images are captured through an uncontaminated glass plate. While this setup introduces no visible obstructions, it causes minor optical effects such as reflections or light attenuation. As shown in the histogram analysis (Figure 4), the pixel-intensity distribution of clean images is highly similar to that of the original case.

Water Droplets. In real-world deployments, lens protectors are often exposed to raindrops or water splashes. To replicate this optical degradation, we sprayed water onto the glass plate, causing each droplet to produce complex refraction and challenging artifacts for both segmentation and restoration.

Condensation. Lens protectors often fog up during transitions from outdoor to indoor or in humid conditions. To replicate this condensation, we exposed the glass plate to hot water steam, creating widespread blur.

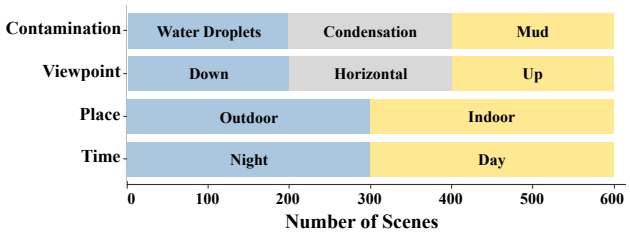

Figure 6. Scene-level statistics. Distribution of contamination types, viewpoints, places, and times across 600 scenes.

Mud. The mud condition simulates severe soiling by splashing a dense wet-soil mixture onto the glass. This causes heavy occlusion of large scene regions and obscures fine details, posing a significant challenge for perception.

Distance-Based Analysis. Figure 5 illustrates the effects of contamination types in different lens-to-protector distances, by presenting the differences in PSNR at each distance with the average PSNR over distances. In the clean case, image distortion primarily arises from light scattering by the protector and monotonically increases with distance. Meanwhile, condensation exhibits a distinct improvement at larger distances, as the layer moves away from the focal plane and becomes defocused. In contrast, for localized occlusions (*e.g.,* mud, water droplets), as distance changes, the occlusion pattern shifts without altering its overall severity.

4. Benchmark

4.1. Dataset Split & Evaluation Protocol

The CLP dataset enables comprehensive benchmarking of semantic segmentation and image restoration tasks under diverse contamination scenarios. It comprises 600 scenes, split into 480 for training and 120 for testing. Figure 6 summarizes scene-level statistics, including contamination types and capture conditions. Each scene contains nine aligned images: one original and a clean–contaminated pair at each of four distances. For evaluation, we analyze the effect of lens-to-glass distance by testing models separately at each distance level. Even with the same type of contamination, visual degradation varies significantly by distance.

4.2. Semantic Segmentation

Every class occurs at least once in the test split, ensuring class-balanced evaluation. Figure 7 summarizes statistics regarding class distributions and frequencies.

Annotation. Direct pixel-wise semantic labels on heavily degraded images are unreliable due to occlusions and distortions. Our controlled capture setup maintained consistent viewpoints across original and degraded image pairs, allowing us to annotate only the *original* (protector-free) images and reuse the same masks for the corresponding degraded images. The CLP dataset comprises 125 semantic classes

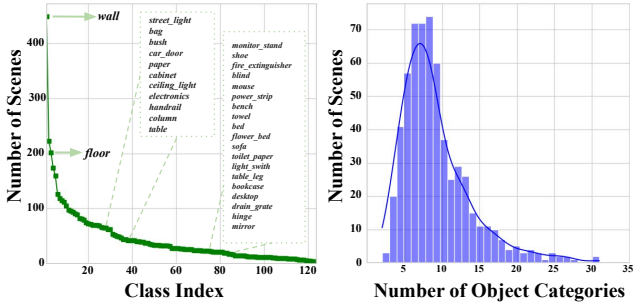

Figure 7. Object distribution statistics. (**Left**) Shareability of object categories sorted by the number of scenes in which each category appears. Common categories (*e.g.,* wall, floor, table) appear in most scenes, while others are limited to a few, reflecting diverse scene compositions. (**Right**) Distribution of object categories per scene.

that cover a broad range of real-world objects and scenes. All masks were manually created and verified through a multi-stage cross-checking process by four expert annotators.

4.3. Image Restoration

Our controlled acquisition setup provides aligned contaminated–original image pairs, enabling supervised learning and reference-based evaluation for image restoration. We adopt the same train/test scene split as in segmentation to maintain consistency across tasks. This unified protocol supports joint experiments, such as evaluating segmentation performance on restored images, and facilitates a comprehensive analysis of how restoration impacts perception under contamination.

5. Experiments

5.1. Experimental Setup

Segmentation Baseline. We evaluated diverse models for semantic segmentation with the following categories:

- **Supervised Learning** provides a reference baseline. Mask2Former [11] and VMamba [44] with full access to semantic annotations across all contamination types and distances in the CLP training set.
- **Domain Generalization (DG)** aims to generalize to unseen contamination types. DG models are trained only on *original* (protector-free) images. The target domain (contaminated images) is entirely unseen during training. We evaluated SoMA [75] and Rein [68], which utilize parameter-efficient adaptation and foundation model integration, respectively.
- **Domain Adaptation (DA)** bridges the gap between the source and target domains using unlabeled target domain data. DAFormer [23] and MIC [24] are trained on original images with semantic labels (source domain) and contaminated images *without* the labels (target domain).
- **Foundation Models** are large-scale vision encoders pretrained on diverse datasets with broad coverage of tasks and domains. Following prior work [70], we use DI-

Table 1. Comparison of semantic segmentation baselines on the CLP test set. Results are reported as mIoU (%).

Method	Clean					Mud					Water					Condensation					Overall
	0 cm	1 cm	3 cm	5 cm	Avg.	0 cm	1 cm	3 cm	5 cm	Avg.	0 cm	1 cm	3 cm	5 cm	Avg.	0 cm	1 cm	3 cm	5 cm	Avg.	
Supervised Learning																					
M2F [11]	43.7	44.7	43.3	44.3	44.0	34.5	27.6	29.9	26.1	29.5	40.3	42.1	38.9	37.8	39.8	27.6	33.6	34.3	34.3	32.5	36.4
VMamba [44]	48.2	47.9	48.1	47.5	47.9	33.0	30.9	28.7	24.7	29.3	40.4	39.5	39.3	38.8	39.5	32.2	34.2	34.2	34.3	33.7	37.6
Domain Generalization																					
Rein [68]	56.9	56.2	56.3	55.6	56.2	38.7	37.3	36.6	33.9	36.6	**56.1**	54.0	52.8	**53.2**	**54.0**	35.8	42.7	40.2	43.4	40.5	46.9
SoMA [75]	**58.5**	**57.9**	**56.9**	**56.5**	**57.4**	44.2	41.0	38.5	36.3	40.0	52.0	51.8	50.8	50.9	51.4	**41.1**	**44.1**	**43.0**	**44.7**	**43.2**	**48.0**
Domain Adaptation																					
DAFormer [23]	45.6	45.1	44.8	44.1	44.9	30.0	28.4	27.0	24.8	27.5	35.9	35.6	34.5	34.5	35.1	26.2	28.7	28.9	30.2	28.5	34.0
MIC [24]	40.0	39.7	39.3	38.8	39.5	29.0	27.3	24.4	23.5	26.1	33.9	33.9	32.5	32.3	33.2	29.0	30.7	28.6	31.6	30.0	32.2
Foundation Models																					
DINOv2 [48]	53.1	52.7	52.9	51.7	52.6	**44.6**	**41.6**	**40.1**	**37.2**	**40.9**	53.5	**54.7**	**53.3**	51.5	53.2	40.6	42.2	42.3	43.1	42.1	47.2
SAM [33]	46.7	45.6	45.8	46.0	46.0	35.7	32.0	27.9	25.1	30.2	47.9	47.4	45.6	43.9	46.2	31.3	34.3	33.4	38.8	34.4	39.2
Robustness Methods																					
RobustSAM [10]	45.6	46.0	45.4	45.1	45.5	38.2	33.6	33.9	28.9	33.6	46.5	45.3	43.0	42.6	44.4	31.2	34.5	34.1	37.9	34.4	39.5
FIFO [36]	38.1	38.2	38.3	38.1	38.2	30.2	27.5	26.8	25.4	27.5	33.7	33.5	33.2	32.3	33.2	27.3	30.6	29.8	32.1	30.0	32.2
Task-aware Restoration																					
UniRestore [6]	42.6	42.4	43.2	42.4	42.7	34.1	31.3	30.4	27.4	30.8	39.5	40.4	39.1	39.9	40.0	32.9	36.3	36.2	37.0	35.6	37.3
URIE [59]	41.1	41.3	40.3	40.4	40.8	27.7	24.2	25.2	25.5	25.6	35.8	36.3	36.0	33.2	35.3	21.6	26.6	28.8	31.8	27.2	32.2

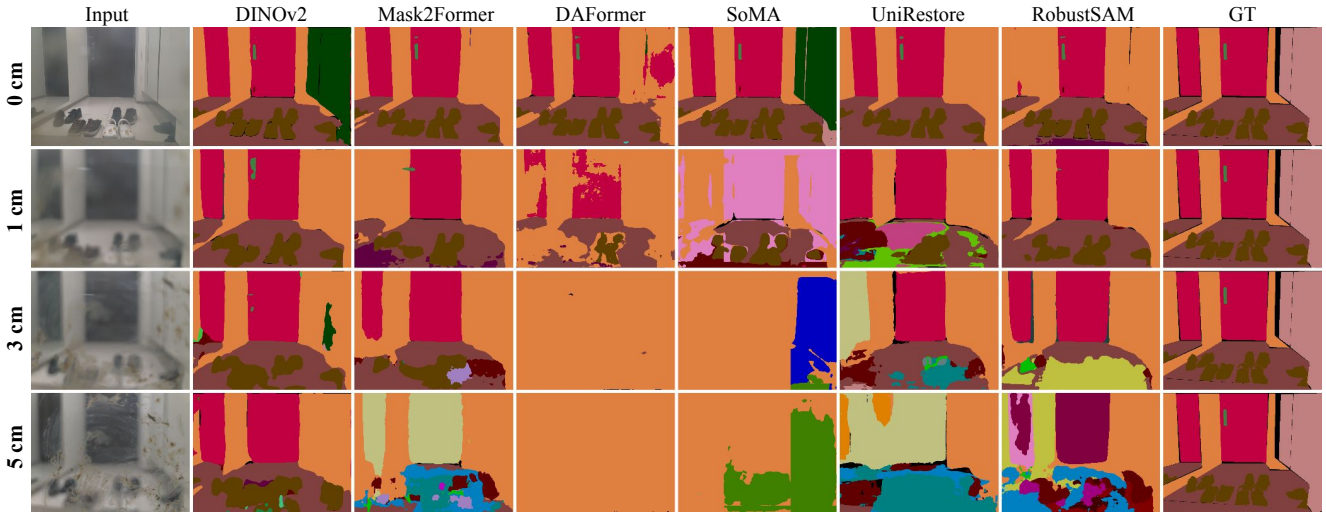

Figure 8. Qualitative comparison of segmentation results under mud contamination across lens-to-protector distances (0cm, 1cm, 3cm, 5cm). DINOv2 remains consistent at 1–5 cm, despite coarse boundaries, but the others degrade severely at 5cm due to heavy blur and occlusion.

NOv2 [48] and SAM [33] as frozen encoders with a Mask2Former decoder trained in a supervised manner.

- **Robustness Methods** are designed to improve segmentation directly under degraded inputs, without explicit image restoration. FIFO [36] learns degradation-invariant features from unlabeled CLP pairs via a fog-pass filter, followed by full segmentation training. We integrate the RobustSAM [10] encoder with a Mask2Former [10] decoder and trained on CLP.
- **Task-aware Restoration** enhances degraded inputs to improve downstream task performance. We apply URIE [59] as a plug-in restoration module with original pretrained weights, following its design that requires no task-specific retraining. UniRestore [6] is trained on CLP in two stages: diffusion-based restoration and task-specific adaptation under the supervision of the frozen Mask2Former [11].

Restoration Baseline. We focus on restoration to support subsequent segmentation tasks, while previous work [13] explores the advancement of restoration. Effective supervised-learning-based restoration models such as NAFNet [7], Restormer [76], MambaIR [19], and DiffUIR [81] serve as our baseline. DiffUIR [81] is a universal restoration model trained on all types of contamination, whereas others have different models for each kind of contamination. Additionally, URIE [59] and UniRestore [6] are included as restoration baselines to evaluate their image enhancement quality.

Table 2. Comparison of restoration baselines on CLP test set. Results are reported as PSNR (dB).

Method	Clean					Mud					Water					Condensation					Overall
	0 cm	1 cm	3 cm	5 cm	Avg.	0 cm	1 cm	3 cm	5 cm	Avg.	0 cm	1 cm	3 cm	5 cm	Avg.	0 cm	1 cm	3 cm	5 cm	Avg.	
NAFNet [7]	30.51	29.39	29.31	29.01	29.55	21.74	21.15	21.54	22.90	21.83	27.42	27.21	27.33	26.99	27.23	23.94	24.28	24.61	25.03	24.46	25.76
Restormer [76]	30.56	29.48	29.32	29.02	29.59	21.03	19.83	20.22	20.96	20.51	26.73	26.51	26.45	26.19	26.47	23.87	24.24	24.36	24.90	24.34	25.22
DiffUIR [81]	**31.91**	**31.31**	**31.23**	**30.61**	**31.26**	23.27	**23.27**	**23.85**	**24.26**	23.66	**30.03**	**29.92**	**29.65**	**29.26**	**29.71**	**26.20**	**27.00**	**27.67**	**27.45**	**27.08**	**27.92**
MambaIR [19]	30.47	29.45	29.38	29.16	29.61	21.35	20.49	21.11	21.70	21.16	26.93	26.82	26.83	26.52	26.77	23.85	24.42	24.43	25.07	24.44	25.49
URIE [59]	24.68	24.13	24.17	23.99	24.24	19.35	18.81	18.99	19.05	19.05	23.24	23.18	23.12	22.98	23.13	20.37	20.85	21.29	21.96	21.26	21.88
UniRestore [6]	30.34	29.70	29.70	29.20	29.73	**23.88**	23.16	23.64	24.04	**23.68**	28.41	28.39	28.49	28.13	28.35	25.40	25.68	25.62	25.82	25.63	26.85

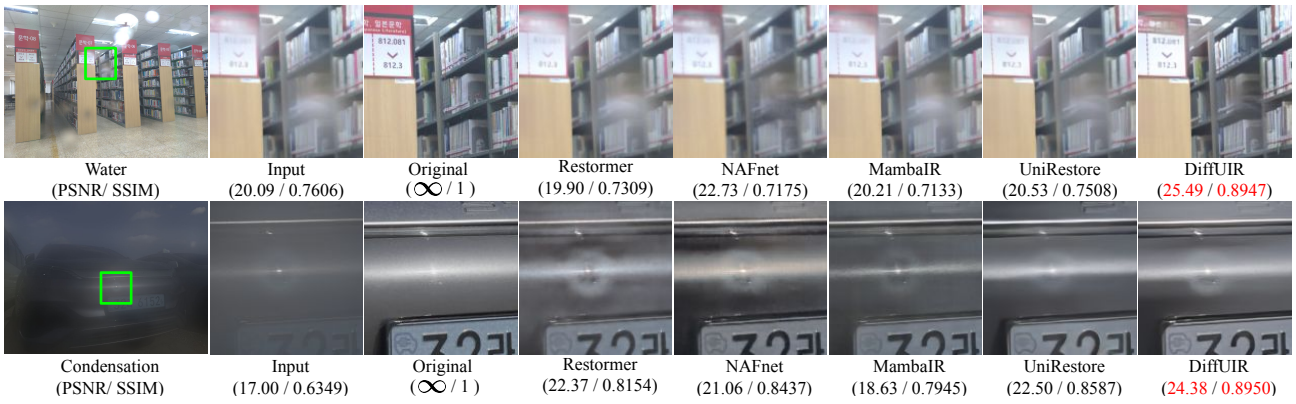

Figure 9. Qualitative comparison of restoration results on Water and Condensation contamination. Each type of contamination generates unique image degradation. DiffUIR [81] shows the best performance, but visible artifacts remain under severe degradation.

Implementation Details. We adopt the official training pipelines provided by each baseline, with minimal modifications for compatibility with our dataset. During training, input images are resized to 512×512 for segmentation and 128×128 for restoration, and we apply each codebase's default augmentations (*e.g.,* random flips and crops); batch sizes are 8 and 4, respectively. During evaluation, segmentation predictions are resized to the training crop size before computing mIoU, while restoration outputs are evaluated at their original resolution using PSNR. All experiments run on a single NVIDIA RTX 4090 GPU with fixed random seeds. Please refer to the Supplement for detailed hyperparameters.

5.2. Semantic Segmentation Results

Table 1 presents the semantic segmentation performance across different contamination types, distances, and learning paradigms. All models exhibit a consistent performance drop under severe mud contamination, especially as the lens-to-protector distance increases. In contrast, other types of contamination show only minor performance changes across distances due to weaker occlusion. Domain Generalization (DG) methods consistently outperform Domain Adaptation (DA) approaches, with SoMA [75] and Rein [68] achieving improvements of +14.0 and +12.9 mIoU over the leading DA baseline. This improvement reveals a key insight: DG methods, trained only on clean images, outperform DA methods despite the latter having access to unlabeled contaminated data. DINOv2 [48] achieves the highest average mIoU (47.2%) across all contamination types, likely due to its large-scale pretraining on diverse data and strong visual prior. In contrast, SAM [33] performs worse despite its scale, reflecting the weaker generalization of its prompt-driven design under distortions and occlusion. Interestingly, robustness-oriented and task-aware methods show mixed results. RobustSAM [10] improves noticeably over SAM [33] under mud (33.6% *vs.* 30.2%), yet its object-centric priors still struggle with non-uniform artifacts. FIFO [36] struggles to generalize to real physical contamination due to its fog-oriented design. Similarly, URIE [59] degrades below the baseline [11], as generic restoration priors are ineffective against such physical artifacts. UniRestore [6] noticeably boosts the baseline under mud and condensation. Further analysis is provided in the joint restoration–segmentation experiments in Section 5.5. A qualitative comparison in Figure 8 shows that contamination at different distances leads to a significant domain shift and substantial degradation. At 0 cm, where contamination effects are minimal, most models produce reasonable segmentation masks for the main scene structures. However, as the distance increases, most models struggle significantly at 5 cm. In contrast, DINOv2 maintains robustness, preserving at least coarse semantic structures (e.g., shoe and door) even under severe distortion at 5 cm. More examples are provided in the Supplement.

Table 3. Comparison of joint restoration-segmentation pipelines on the CLP test set. Results are reported as mIoU (%). We report contamination-wise averages and the overall average.

Method	Water	Mud	Cond.	Overall
Mask2Former	39.8	29.5	32.5	33.9
NAFNet + Mask2Former	35.2 (-4.6)	25.3 (-4.2)	32.2 (-0.3)	33.7 (-2.7)
DiffUIR + Mask2Former	37.9 (-1.9)	29.2 (-0.3)	34.9 (+2.4)	35.7 (-0.7)
URIE + Mask2Former	35.3 (-4.5)	25.6 (-3.9)	27.2 (-5.3)	32.2 (-4.2)
UniRestore + Mask2Former	40.0 (+0.2)	30.8 (+1.3)	35.6 (+3.1)	35.5 (+1.6)

Table 4. Segmentation (mIoU/pAcc) and restoration (PSNR/SSIM) performance vs. CLP training set ratio (10–100%).

	Scenes for Training				
Method	10%	25%	50%	75%	100%
Semantic Segmentation (mIoU / pAcc)					
DAFormer	11.56 / 47.94	20.84 / 58.28	28.50 / 65.71	32.64 / 69.16	**34.73 / 71.59**
SoMA	22.19 / 60.56	34.75 / 68.74	41.64 / 72.83	46.68 / 75.47	**48.69 / 76.92**
Image Restoration (PSNR / SSIM)					
Restormer	24.54 / 0.7802	24.87 / 0.7957	25.00 / 0.7975	25.16 / 0.8015	**25.22 / 0.8040**
NAFNet	24.63 / 0.7925	25.28 / 0.8042	25.39 / 0.8088	25.50 / 0.8105	**25.77 / 0.8147**

5.3. Image Restoration Results

Table 2 and Figure 9 present the quantitative and qualitative comparisons of restoration models on our dataset. DiffUIR [81] and UniRestore [6] achieve the highest overall PSNRs of 27.92 dB and 26.85 dB, respectively. Both diffusion-based methods outperform supervised baselines such as NAFNet [7] (25.76 dB) and Restormer [76] (25.22 dB), indicating that generative priors are effective for the complex degradation patterns in CLP. However, artifacts often remain in severely degraded regions after restoration (see Figure 9). This suggests that restoration under severe real-world contamination remains an open challenge.

5.4. Model Complexity Analysis

Figure 10 shows the trade-offs between performance and model complexity for segmentation and restoration. For segmentation (**left**), relatively efficient models [48, 68, 75] achieve the best overall mIoU while requiring substantially less computation than larger models such as SAM [33] and RobustSAM [10]. This indicates that robustness under contamination depends more on adaptation strategy than on model scale. UniRestore [6] incurs the highest cost (2512 G) but achieves only mid-tier segmentation accuracy (37.3 mIoU). It further confirms that naïve scaling offers no advantage. For restoration (**right**), DiffUIR [81] achieves the best PSNR (27.92 dB) at minimal cost (33 G). The remaining deterministic methods cluster within a narrow 0.5 dB band regardless of their complexity. This suggests that generative approaches (*e.g.*, diffusion-based) are a more promising direction than further scaling deterministic models.

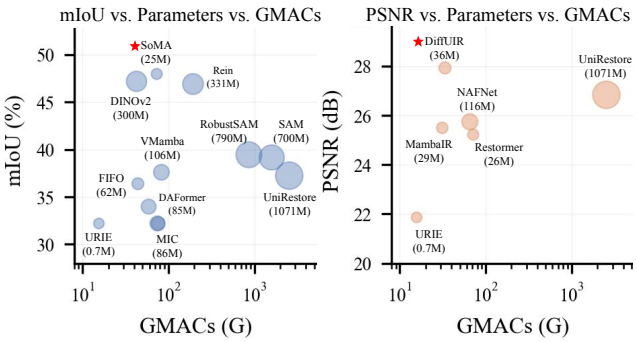

Figure 10. Model complexity *vs.* performance for segmentation (**left**) and restoration (**right**). Segmentation models show a trade-off between accuracy and complexity, while restoration models exhibit no consistent pattern.

5.5. Ablation Study

Data Scale. Table 4 presents the effect of training set size (10%–100%), with results averaged over all contamination types in the test set. Increasing the training data generally improves both segmentation and restoration accuracy, though gains become marginal at larger scales. Exploring further data scaling remains an interesting direction for future work.

Joint Restoration-Segmentation Pipeline. We further investigate how restoration affects the downstream segmentation by combining restoration models with Mask2Former [11]. Table 3 shows that higher visual quality does not strictly guarantee better semantic segmentation. DiffUIR [81] achieves the best restoration performance in Table 2. However, its benefit for segmentation remains limited, improving the baseline under condensation (+2.4) while slightly reducing performance on water (-1.9) and mud (-0.3). UniRestore [6] exhibits a notable advancement. As one can see, it is the only method that consistently boosts the baseline (+1.6 mIoU overall). In particular, it achieves significant gains under condensation (+3.1) and mud (+1.3). These observations suggest that task-aware restoration is a promising direction for perception under real-world contamination.

6. Conclusion

We introduce CLP (Contaminated Lens Protectors), a real-world dataset for benchmarking robust perception under realistic sensor contamination. CLP captures multiple types of lens-protector degradations—mud, water droplets, and condensation—across varied lens-to-protector distances. It provides paired clean and contaminated images, dense semantic masks, and aligned restoration targets. Our experiments benchmark diverse segmentation and restoration approaches, including data scale effects and joint restoration–segmentation pipelines. These results establish baselines and insights for robust perception under real-world contaminated-vision scenarios.

Acknowledgements

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the Advanced GPU Utilization Support Program, the National Program for Excellence in SW (2024-0-00071), and the graduate school support program (IITP-2026-RS-2024-00430997, IITP-2026-RS-2024-00426853) supervised by the IITP (Institute of Information & communications Technology Planning & evaluation) and the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (RS-2025-24803335) and the Ministry of Education (MOE) for the G-LAMP program (RS-2025-25441317), and Seoul Metropolitan Government for the RISE program (2025-RISE-01-020-04), and Basic Science Research Program (RS-2021-NR060140). We thank Corca AI for sponsoring the compute used in this work.

References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S. Brown. A high-quality denoising dataset for smartphone cameras. In *CVPR*, 2018. 3

[2] Cosmin Ancuti, Codruta O Ancuti, Radu Timofte, and Christophe De Vleeschouwer. I-haze: A dehazing benchmark with real hazy and haze-free indoor images. In *ACIVS*, 2018. 3

[3] Codruta O Ancuti, Cosmin Ancuti, Mateu Sbert, and Radu Timofte. Dense-haze: A benchmark for image dehazing with dense-haze and haze-free images. In *ICIP*, 2019. 3

[4] Yasser Benigmim, Subhankar Roy, Slim Essid, Vicky Kalogeiton, and Stéphane Lathuilière. Collaborating foundation models for domain generalized semantic segmentation. In *CVPR*, 2024. 3

[5] Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *CVPR*, 2020. 1, 3

[6] I-Hsiang Chen, Wei-Ting Chen, Yu-Wei Liu, Yuan-Chun Chiang, Sy-Yen Kuo, and Ming-Hsuan Yang. Unirestore: Unified perceptual and task-oriented image restoration model using diffusion prior. In *CVPR*, 2025. 3, 6, 7, 8

[7] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *ECCV*, 2022. 3, 6, 7, 8

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 40(4):834–848, 2017. 3

[9] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 3

[10] Wei-Ting Chen, Yu-Jiet Vong, Sy-Yen Kuo, Sizhou Ma, and Jian Wang. Robustsam: Segment anything robustly on degraded images. In *CVPR*, 2024. 3, 6, 7, 8

[11] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 3, 5, 6, 7, 8

[12] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *CVPR*, 2021. 3

[13] Sooyoung Choi, Sungyong Park, and Heewon Kim. Sidl: A real-world dataset for restoring smartphone images with dirty lenses. In *AAAI*, 2025. 1, 3, 6

[14] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 3

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[16] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, 2019. 3

[17] Jinwei Gu, Ravi Ramamoorthi, Peter Belhumeur, and Shree Nayar. Dirty glass: rendering contamination on transparent surfaces. In *EGSR*, 2007. 3

[18] Jinwei Gu, Ravi Ramamoorthi, Peter Belhumeur, and Shree Nayar. Removing image artifacts due to dirty camera lenses and thin occluders. *ACM Trans. Graph.*, 28, 2009. 3

[19] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. In *ECCV*, 2024. 3, 6, 7

[20] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 3

[21] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Task-driven super resolution: Object detection in low-resolution images. *TPAMI*, 44(11):7896–7914, 2021. 3

[22] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, 2018. 3

[23] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *CVPR*, 2022. 5, 6

[24] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. Mic: Masked image consistency for context-enhanced domain adaptation. In *CVPR*, 2023. 3, 5, 6

[25] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, and Pheng-Ann Heng. Depth-attentional features for single-image rain removal. In *CVPR*, 2019. 1, 3

[26] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, 2019. 3

[27] Christoph Kamann and Carsten Rother. Benchmarking the robustness of semantic segmentation models. In *CVPR*, 2020. 1, 3

[28] Christoph Kamann and Carsten Rother. Benchmarking the robustness of semantic segmentation models with respect to common corruptions. *IJCV*, 2021. 1

[29] Heewon Kim, Myungsub Choi, Bee Lim, and Kyoung Mu Lee. Task-aware image downscaling. In *ECCV*, 2018. 3

[30] Jin Kim, Jiyoung Lee, Jungin Park, Dongbo Min, and Kwanghoon Sohn. Pin the memory: Learning to generalize semantic segmentation. In *CVPR*, 2022. 3

[31] Jaeha Kim, Junghun Oh, and Kyoung Mu Lee. Beyond image super-resolution for image recognition with task-driven perceptual loss. In *CVPR*, 2024. 3

[32] Jaeha Kim, Junghun Oh, and Kyoung Mu Lee. Exploiting diffusion prior for task-driven image restoration. In *ICCV*, 2025. 3

[33] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 1, 3, 6, 7, 8

[34] Lingshun Kong, Jiangxin Dong, Jianjun Ge, Mingqiang Li, and Jinshan Pan. Efficient frequency domain-based transformers for high-quality image deblurring. In *CVPR*, 2023. 3

[35] Suhyeon Lee, Hongje Seong, Seongwon Lee, and Euntai Kim. Wildnet: Learning domain generalized semantic segmentation from the wild. In *CVPR*, 2022. 3

[36] Sohyun Lee, Taeyoung Son, and Suha Kwak. Fifo: Learning fog-invariant features for foggy scene segmentation. In *CVPR*, 2022. 3, 6, 7

[37] Sohyun Lee, Yeho Gwon, Lukas Hoyer, and Suha Kwak. GaRA-SAM: Robustifying segment anything model with gated-rank adaptation. In *NeruIPS*, 2025. 3

[38] Siyuan Li, Iago Breno Araujo, Wenqi Ren, Zhangyang Wang, Eric K Tokuda, Roberto Hirata Junior, Roberto Cesar-Junior, Jiawan Zhang, Xiaojie Guo, and Xiaochun Cao. Single image deraining: A comprehensive benchmark analysis. In *CVPR*, 2019. 1, 3

[39] Xiaoyu Li, Bo Zhang, Jing Liao, and Pedro V Sander. Let's see clearly: Contaminant artifact removal for moving cameras. In *ICCV*, 2021. 3

[40] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 3

[41] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *ECCV*, 2024. 3

[42] Ding Liu, Bihan Wen, Jianbo Jiao, Xianming Liu, Zhangyang Wang, and Thomas S Huang. Connecting image denoising and high-level vision tasks via deep learning. *TIP*, 29:3695–3706, 2020. 3

[43] Wenyu Liu, Gaofeng Ren, Runsheng Yu, Shi Guo, Jianke Zhu, and Lei Zhang. Image-adaptive yolo for object detection in adverse weather conditions. In *AAAI*, 2022. 3

[44] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. VMamba: Visual state space model. In *NeruIPS*, 2024. 3, 5, 6

[45] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 3

[46] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017. 3

[47] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017. 3

[48] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3, 6, 7, 8

[49] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *CVPR*, 2020. 3

[50] Tobias Plotz and Stefan Roth. Benchmarking denoising algorithms with real photographs. In *CVPR*, 2017. 3

[51] Horia Porav, Tom Bruls, and Paul Newman. I can see clearly now: Image restoration via de-raining. In *ICRA*, 2019. 3

[52] Horia Porav, Valentina-Nicoleta Musat, Tom Bruls, and Paul Newman. Rainy screens: Collecting rainy datasets, indoors. *arXiv preprint arXiv:2003.04742*, 2020. 3

[53] AN Rajagopalan et al. Improving robustness of semantic segmentation to motion-blur using class-centric augmentation. In *CVPR*, 2023. 3

[54] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 3

[55] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *ECCV*, 2020. 3

[56] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *IJCV*, 2018. 3

[57] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *ICCV*, 2019. 1, 3

[58] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *CVPR*, 2021. 1, 3

[59] Taeyoung Son, Juwon Kang, Namyup Kim, Sunghyun Cho, and Suha Kwak. Urie: Universal image enhancement for visual recognition in the wild. In *ECCV*, 2020. 3, 6, 7

[60] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *ICCV*, 2021. 3

[61] Shangquan Sun, Wenqi Li, Jingzhi Zhang, and Lei Zhang. Rethinking image restoration for object detection. In *NeurIPS*, 2022. 3

[62] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, 2018. 3

[63] Michal Uricar, Ganesh Sistu, Hazem Rashed, Antonin Vobecky, Varun Ravi Kumar, Pavel Krizek, Fabian Burger, and Senthil Yogamani. Let's get dirty: Gan based data augmentation for camera lens soiling detection in autonomous driving. In *WACV*, 2021. 3

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3

[65] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Dada: Depth-aware domain adaptation in semantic segmentation. In *ICCV*, 2019. 3

[66] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 3

[67] Yufei Wang, Renjie Wan, Wenhan Yang, Bihan Wen, Lappui Chau, and Alex C Kot. Removing image artifacts from scratched lens protectors. *arXiv*, 2023. 1, 3

[68] Zhixiang Wei, Lin Chen, Yi Jin, Xiaoxiao Ma, Tianle Liu, Pengyang Ling, Ben Wang, Huaian Chen, and Jinjin Zheng. Stronger fewer & superior: Harnessing vision foundation models for domain generalized semantic segmentation. In *CVPR*, 2024. 3, 5, 6, 7, 8

[69] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS*, 2021. 3

[70] Xin Yang, Xin Zhang, and Xinchao Wang. Erf: A benchmark dataset for robust semantic segmentation under extreme rainfall conditions. In *AAAI*, 2025. 5

[71] Tian Ye, Sixiang Chen, Wenhao Chai, Zhaohu Xing, Jing Qin, Ge Lin, and Lei Zhu. Learning diffusion texture priors for image restoration. In *CVPR*, 2024. 3

[72] Senthil Yogamani, Ciarán Hughes, Jonathan Horgan, Ganesh Sistu, Padraig Varley, Derek O'Dea, Michal Uricár, Stefan Milz, Martin Simon, Karl Amende, et al. Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving. In *ICCV*, 2019. 3

[73] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 3

[74] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020. 3

[75] Seokju Yun, Seunghye Chae, Dongheon Lee, and Youngmin Ro. Soma: Singular value decomposed minor components adaptation for domain generalizable representation learning. In *CVPR*, 2025. 3, 5, 6, 7, 8

[76] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022. 3, 6, 7, 8

[77] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *TIP*, 26(7):3142–3155, 2017. 3

[78] Yuhong Zhang, Hengsheng Zhang, Xinning Chai, Zhengxue Cheng, Rong Xie, Li Song, and Wenjun Zhang. Diff-restorer: Unleashing visual prompts for diffusion-based universal image restoration. *arXiv preprint arXiv:2407.03636*, 2024. 3

[79] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 3

[80] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Point-wise spatial attention network for scene parsing. In *ECCV*, 2018. 3

[81] Dian Zheng, Xiao-Ming Wu, Shuzhou Yang, Jian Zhang, Jian-Fang Hu, and Wei-Shi Zheng. Selective hourglass mapping for universal image restoration based on diffusion model. In *CVPR*, 2024. 3, 6, 7, 8

[82] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, 2021. 3

[83] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 2019. 1, 3