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

Sungyong Park[1], Sooyoung Choi[1], Hyunseo Koh[1], Youngjae Choi[1], Heewon Kim[2,3]

[1]Soongsil University

[2]School of AI·Software, College of AI, Soongsil University [3]Kairoba Inc.

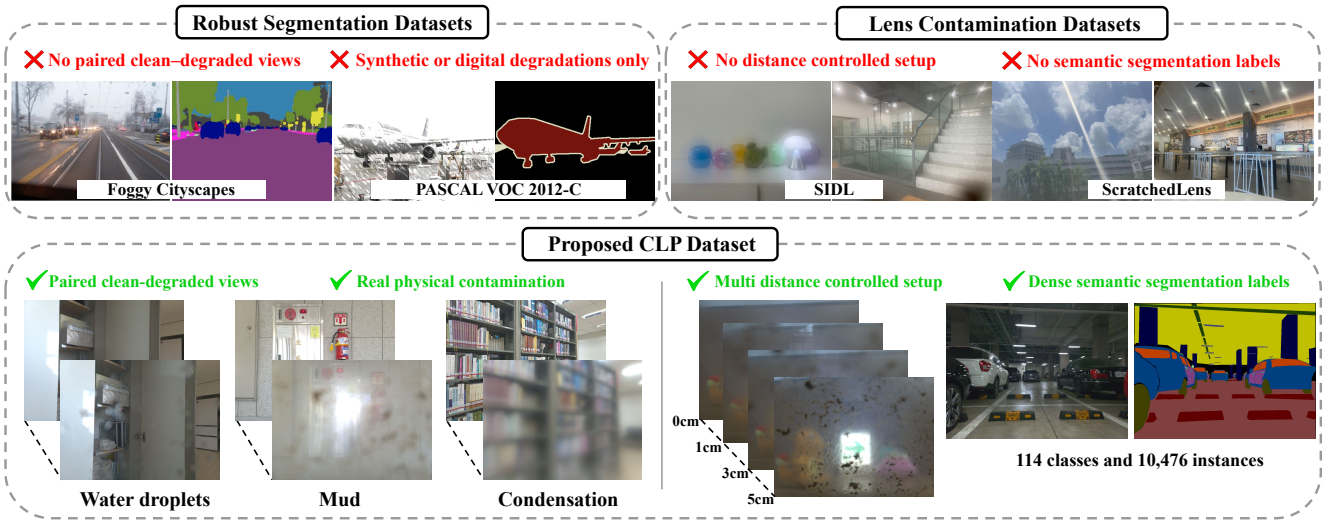


Figure 1. Existing robustness and lens contamination datasets have a limited scope, lacking realistic contamination, aligned pairs, or annotations. CLP offers real physical contamination with paired clean/degraded image pairs, distance control, and dense semantic labels.

## 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. This dataset 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.*

## 1. Introduction

Semantic segmentation has achieved remarkable progress with the advent of large-scale datasets [25, 54] and foundational models [20]. This progress has enabled high-precision scene understanding by capturing rich spatial and semantic cues. For robust perception, recent benchmarks cover digital corruptions [17, 18] and environmental conditions like adverse weather [5, 16, 23, 37, 38]. 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 [9, 43]. However, these approaches overlook a crucial aspect of real-world lens setups—the physical gap be-

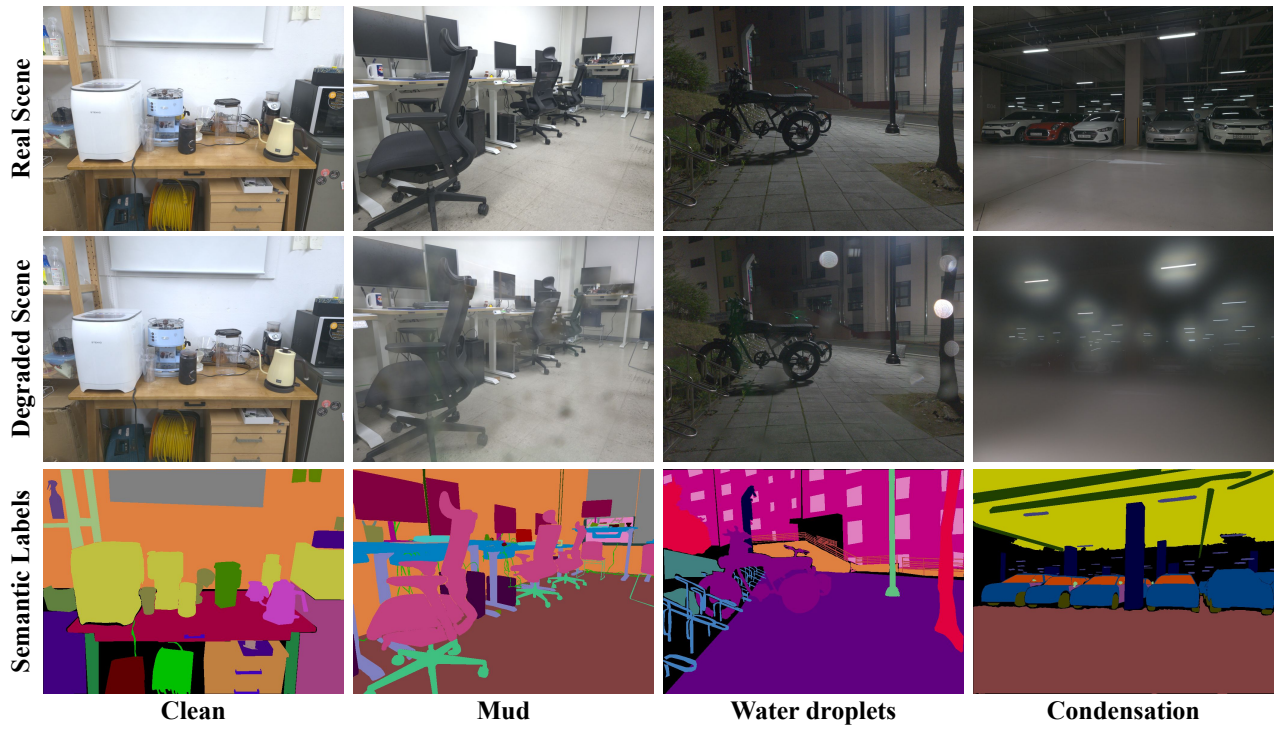


Figure 2. Example scenes from the CLP dataset illustrating lens protector contamination. Each column shows a scene under a different condition: Clean, Mud, Water Droplets, and Condensation. From top to bottom: (1) original (protector-free) scene, (2) degraded version with lens contamination, and (3)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

tween 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 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 **114** object categories with **10,476** 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 114 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 seg-


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.

mentation and restoration models, establishing baselines and providing insights for robust perception systems.

## 2. Related Works

**Generic Semantic Segmentation Methods.**

**Robust Semantic Segmentation Methods.** Current methods for robust segmentation primarily address visual degradations through two principal approaches. A major line of work treats this challenge as a domain shift problem. Domain Adaptation (DA) [13–15, 31, 39, 41] aligns clean and degraded domains, and Domain Generalization (DG) [4, 8, 19, 22, 44, 49] learns domain-invariant representations from multiple sources. These paradigms are well-suited for lens contamination, which presents unique domain shifts characterized by optical distortion and partial occlusions. Another direction leverages foundation models [20, 30, 34, 42] trained on diverse visual data to improve robustness to common degradations. In this work, we leverage CLP to provide a principled evaluation of robustness methods under real lens contamination, outlining their capabilities and limitations to inform future research.

**Robust Semantic Segmentation Datasets.** Robust semantic segmentation aims to ensure reliable scene understanding under various visual degradations such as weather, lighting, and sensor noise. While large-scale datasets [10, 12, 25, 29, 47, 54] have driven significant progress under clean conditions, models still suffer severe performance degradation in adverse environments. Real-world robustness datasets [5, 37, 38, 48] capture complex conditions such as fog, rain, and night scenes; however, they mostly focus on outdoor scenarios and do not provide pixel-level alignment between clear and degraded views. This limitation has led to approaches that synthesize artifacts on clean images, producing scalable aligned pairs for weather effects [16, 23, 36], and digital corruption [17, 33]. However, these methods rely on simulated artifacts that fail to reproduce the complex physical characteristics of real contamination. Prior work [40] explores this approach to lens soiling, yet the generated artifacts remain unrealistic and limited in reproducing real optical effects (*e.g.,* occlusion, blur, scattering). CLP aims to fill this gap by offering real-world lens contamination with paired clean–degraded images and dense segmentation labels, providing a valuable resource for robustness evaluation.

### 2.1. Semantic Segmentation

**Generic Semantic Segmentation.**

**Robust Semantic Segmentation.**

### 2.2. Image Restoration

**Task-Aware Image Restoration.**

**Perceptual Image Restoration.**

### 2.3. Image Restoration.

Image Restoration aims to recover clean images from degraded inputs. Deep learning methods based on CNNs [6, 51] and Transformers [21, 50] have shown strong performance across various restoration tasks. Recently, diffusion-based models [26, 46] and efficient state-space models such as VMamba [11] have further improved performance and flexibility across diverse degradation types, alongside unified all-in-one architectures [52, 53]. To support these advances, several datasets have been proposed to benchmark restoration performance under specific degradations such as noise [1, 32], blur [28, 35], and adverse weather [2, 3]. However, real-world degradations like lens contamination remain largely unexplored. While a few datasets [9, 24, 43] address physical contaminants, they are limited in scope. For instance, recent efforts [9] are limited to smartphones, whose fixed physical configuration prevents systematic variation of

key factors like lens-to-protector distance. In contrast, CLP captures realistic lens-protector contamination across multiple distances and diverse environments (*e.g.,* indoor/outdoor, day/night). By providing aligned clean-degraded pairs with semantic labels, CLP enables joint evaluation of restoration and segmentation robustness.

## 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 a smartphone[1] camera 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.

### 3.3. Preprocessing

All images were captured in raw format ($4000 \times 3000$ 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 $1000 \times 750$ 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.

[1]Samsung Galaxy S20 Ultra

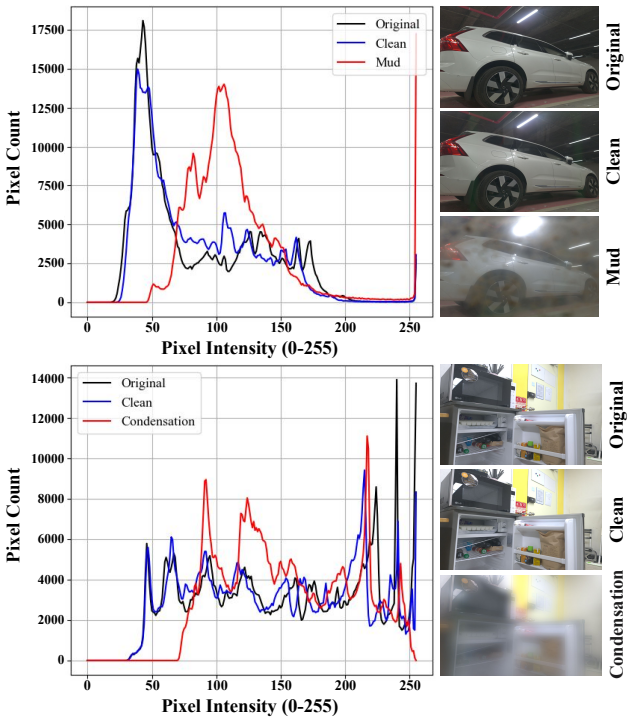


Figure 4. Histogram analysis of contaminated lens protectors. Each plot compares pixel intensity distributions for clean and degraded images (Mud, Condensation). Sample scenes on the right show visual degradation under each condition. Protector-induced contamination reduces contrast and alters pixel distributions, highlighting the need for restoration.

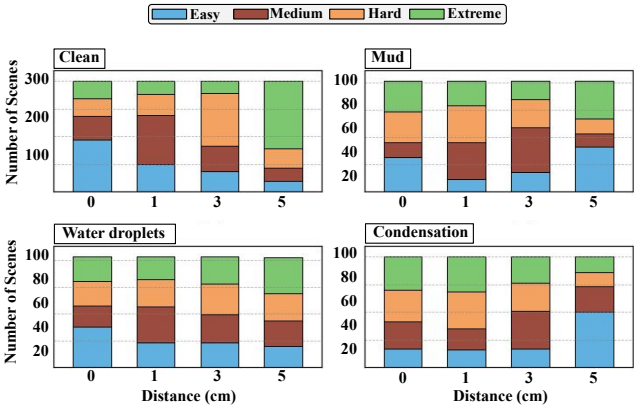


Figure 5. Difficulty distribution across lens-to-protector distances under different contamination types.

**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

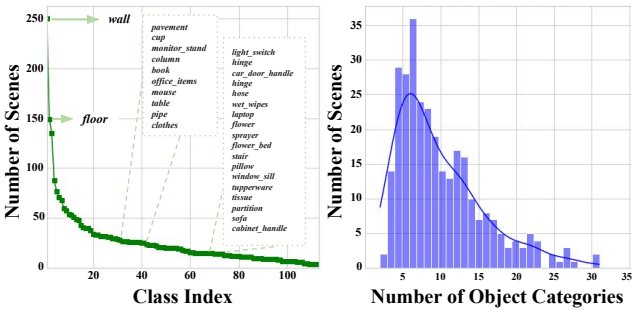


Figure 6. Object distribution statistics. (**Left**) Shareability of object categories sorted by the number of scenes in which each category appears. Common categories (*e.g.,* wall, floor, table) appear in most scenes, while others are limited to a few, reflecting diverse scene compositions. (**Right**) Distribution of object categories per scene. Most scenes contain between 5 and 15 categories, providing sufficiently rich contextual diversity to support recognition tasks.

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.

**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.

**Contamination Intensity by Distance.** To better understand how contamination affects visual quality across different distances, we analyzed the relative PSNR values within each contamination type (Figure 5). Our analysis reveals that spatial occlusion patterns often dominate over distance effects, particularly for mud and water droplets, while condensation shows more uniform degradation regardless of distance. This suggests that the position and coverage of contamination are often more critical than optical distance for real-world lens protection scenarios.

# 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 300 scenes, split into 240 for training and 60 for testing. Figure 7 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

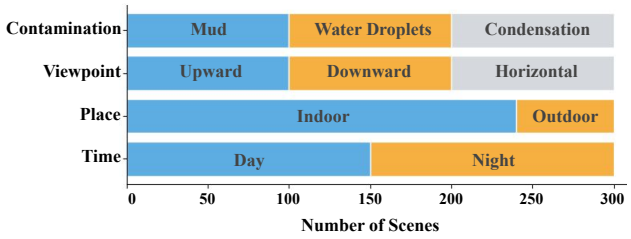


Figure 7. Scene-level statistics. Distribution of contamination types, viewpoints, places, and times across 300 scenes.

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 6 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 114 semantic classes 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 [7] and VMamba [27] 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

Table 1. Comparison of semantic segmentation baselines on the CLP test set. Results are reported as mIoU / pAcc (%). **Bold**: best, underline: second-best *per contamination type*.

| Method | 0 cm | | | | 1 cm | | | | 3 cm | | | | 5 cm | | | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Clean | Water | Mud | Cond | Clean | Water | Mud | Cond | Clean | Water | Mud | Cond | Clean | Water | Mud | Cond | mIoU | pAcc |
| Mask2Former | 41.1/69.5 | 39.5/69.9 | 27.4/68.5 | 30.3/69.2 | 40.2/69.3 | 39.4/69.7 | 23.8/60.5 | 33.6/71.4 | 40.6/70.4 | 37.4/68.7 | 19.3/50.9 | 33.5/69.2 | 40.3/69.7 | 37.4/67.0 | 17.3/52.8 | 32.5/70.6 | 33.4 | 71.1 |
| VMamba | 30.9/67.2 | 24.1/67.0 | 10.4/56.0 | 16.8/62.0 | 31.4/67.3 | 24.0/68.3 | 9.8/51.5 | 16.2/61.8 | 31.6/67.6 | 23.6/67.7 | 9.1/45.1 | 16.9/60.1 | 31.4/67.2 | 23.7/67.2 | 6.7/41.5 | 17.9/62.7 | 21.6 | 65.2 |
| Rein | 52.6/81.8 | 48.3/76.9 | 35.8/71.7 | 30.4/74.0 | 53.2/81.2 | 48.4/79.1 | 33.9/69.5 | 33.5/76.0 | 53.6/82.0 | 49.4/78.2 | 26.5/58.3 | 35.8/75.5 | 52.3/81.8 | 49.1/76.7 | 23.9/53.9 | 33.5/73.7 | 38.2 | 74.4 |
| SoMA | **57.4/82.9** | 55.2/80.3 | 36.1/72.1 | 32.7/74.3 | **57.6/82.4** | 51.9/80.5 | 33.5/70.3 | 36.9/79.8 | **57.2/82.2** | 51.8/77.7 | 25.2/62.0 | 37.2/79.1 | **57.5/82.5** | 51.5/77.1 | 22.5/53.5 | 35.5/**79.4** | 43.7 | 76.0 |
| DAFormer | 45.2/76.8 | 35.4/73.9 | 20.7/67.6 | 22.6/68.8 | 44.9/76.8 | 34.6/74.3 | 18.2/60.9 | 24.5/70.1 | 44.6/76.7 | 34.4/73.4 | 15.8/58.1 | 28.0/71.7 | 43.8/76.0 | 33.7/71.7 | 13.5/53.3 | 26.8/70.7 | 27.9 | 70.0 |
| MIC | 36.1/72.6 | 30.3/73.0 | 19.2/63.1 | 24.7/73.2 | 35.4/72.3 | 30.8/73.3 | 17.6/60.5 | 24.4/71.7 | 35.5/72.3 | 29.5/71.4 | 14.9/55.2 | 25.9/74.4 | 35.4/72.1 | 30.3/69.9 | 12.2/50.4 | 25.9/74.0 | 26.7 | 68.7 |
| DINOv2 | 53.3/82.5 | **56.0/84.0** | **40.2/78.9** | **40.3/78.1** | 52.9/82.2 | **53.8/80.5** | **35.9/71.5** | **42.3/81.8** | 52.4/82.0 | **54.5/80.3** | **31.7/68.5** | **42.4/79.3** | 53.2/82.2 | **55.0/81.6** | **24.2/66.3** | **39.3**/78.9 | **45.4** | **78.6** |
| SAM | 35.2/69.1 | 29.6/65.0 | 26.0/70.4 | 21.7/58.6 | 34.7/69.1 | 29.8/65.3 | 18.8/61.8 | 23.3/58.2 | 34.8/69.8 | 31.8/67.9 | 15.2/57.8 | 26.4/63.5 | 34.8/69.8 | 28.2/63.8 | 14.6/50.0 | 27.4/63.8 | 27.0 | 64.0 |
| URIE+M2F | 35.2/69.1 | 29.6/65.0 | 26.0/70.4 | 21.7/58.6 | 34.7/69.1 | 29.8/65.3 | 18.8/61.8 | 23.3/58.2 | 34.8/69.8 | 31.8/67.9 | 15.2/57.8 | 26.4/63.5 | 34.8/69.8 | 28.2/63.8 | 14.6/50.0 | 27.4/63.8 | 27.0 | 64.0 |
| Un | 35.2/69.1 | 29.6/65.0 | 26.0/70.4 | 21.7/58.6 | 34.7/69.1 | 29.8/65.3 | 18.8/61.8 | 23.3/58.2 | 34.8/69.8 | 31.8/67.9 | 15.2/57.8 | 26.4/63.5 | 34.8/69.8 | 28.2/63.8 | 14.6/50.0 | 27.4/63.8 | 27.0 | 64.0 |

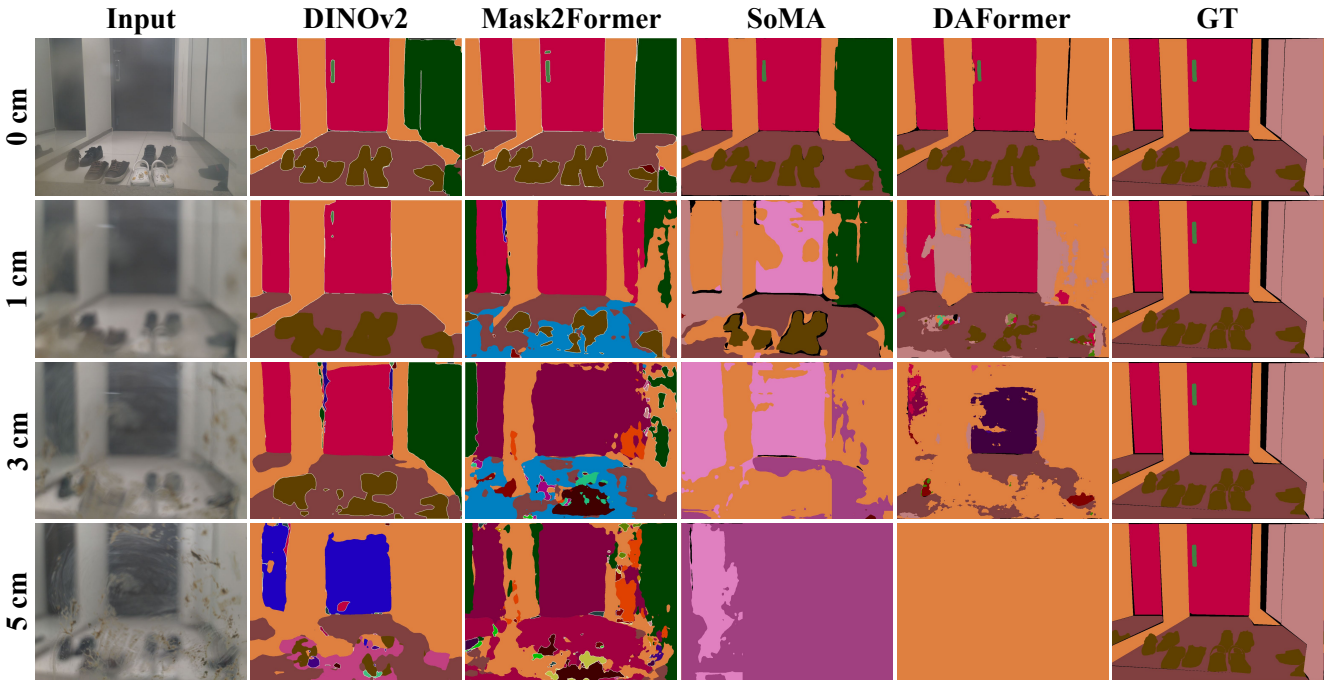


Figure 8. Qualitative comparison of segmentation results under mud contamination across lens-to-protector distances (0cm, 1cm, 3cm, 5cm). DINOv2 remains consistent at 1–3 cm, despite coarse boundaries, but all models degrade severely at 5cm due to heavy blur and occlusion.

(contaminated images) is entirely unseen during training. We evaluated SoMA [49] and Rein [44], 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 [14] and MIC [15] are trained on original images with semantic labels (source domain) and contaminated images *without* the labels (target domain) for adaptation.
- **Foundation Models** are large-scale vision encoders pretrained on diverse datasets with broad coverage of tasks and domains. Following prior work [45], we use DINOv2 [30] and SAM [20] as frozen encoders with a Mask2Former decoder trained in a supervised manner.

**Restoration Baseline.** We focus on restoration to support subsequent segmentation tasks, while previous work [9] explores the advancement of restoration. Effective supervised-learning-based restoration models such as NAFNet [6], Restormer [50], MambaIR [11], and DiffUIR [53] serve as our baseline. DiffUIR [53] is a universal restoration model trained on all types of contamination, whereas others have different models for each kind of contamination.

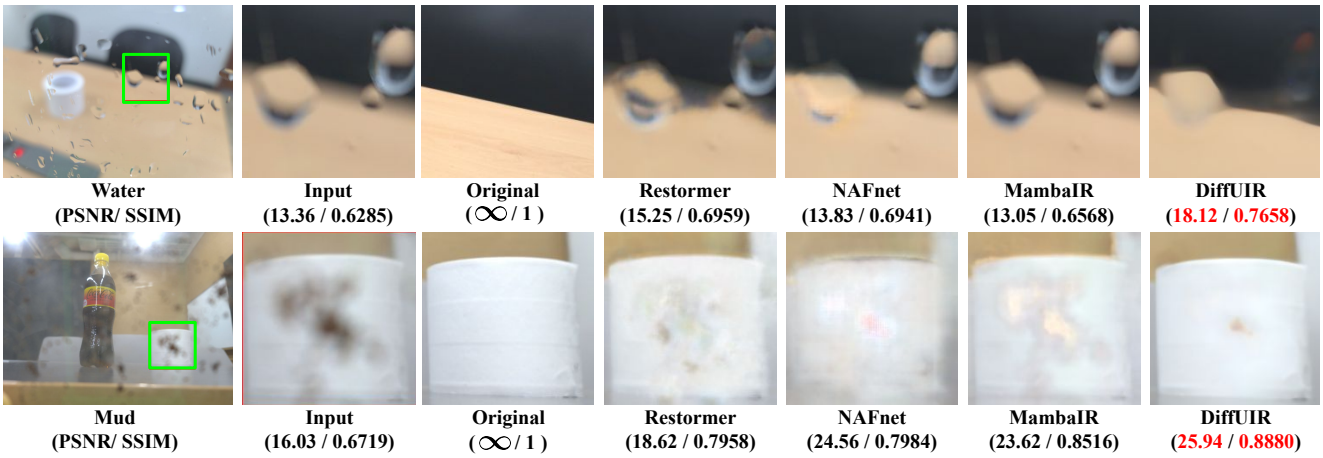


Figure 9. Qualitative comparison of restoration results on Water and Mud contamination. Each type of contamination generates unique image degradation. DiffUIR [53] shows the best performance, but visible artifacts remain under severe degradation.

Table 2. Comparison of restoration baselines on CLP test set. Results are reported as PSNR / SSIM.

| Method | Dist | Clean | Water | Mud | Condensation | Average |
|---|---|---|---|---|---|---|
| | | PSNR/SSIM | PSNR/SSIM | PSNR/SSIM | PSNR/SSIM | PSNR/SSIM |
| NAFNet | 0 cm | 30.23 / 0.8706 | 28.40 / 0.8216 | 22.50 / 0.7401 | 24.58 / 0.8013 | 26.43 / 0.8084 |
| | 1 cm | 29.29 / 0.8615 | 28.35 / 0.8247 | 21.43 / 0.7246 | 24.95 / 0.8064 | 26.01 / 0.8043 |
| | 3 cm | 29.15 / 0.8544 | 27.91 / 0.8093 | 21.90 / 0.7240 | 25.68 / 0.8191 | 26.16 / 0.8017 |
| | 5 cm | 28.30 / 0.8381 | 27.62 / 0.7996 | 22.21 / 0.7240 | 25.60 / 0.8062 | 25.93 / 0.7920 |
| Restormer | 0 cm | 31.18 / 0.8800 | 28.58 / 0.8289 | 21.07 / 0.7320 | 24.54 / 0.7900 | 26.34 / 0.8077 |
| | 1 cm | 30.14 / 0.8687 | 28.61 / 0.8298 | 20.27 / 0.7134 | 25.11 / 0.7911 | 26.03 / 0.8008 |
| | 3 cm | 30.03 / 0.8621 | 28.38 / 0.8221 | 20.84 / 0.7166 | 25.20 / 0.8018 | 26.11 / 0.8007 |
| | 5 cm | 28.93 / 0.8431 | 28.31 / 0.8107 | 21.27 / 0.7105 | 25.09 / 0.7904 | 25.90 / 0.7887 |
| DiffUIR | 0 cm | **33.38 / 0.9146** | **31.81 / 0.8951** | **24.92 / 0.8489** | **27.14 / 0.8671** | **29.31 / 0.8814** |
| | 1 cm | **32.73 / 0.9160** | **31.94 / 0.8993** | **23.97 / 0.8368** | **27.87 / 0.8689** | **29.13 / 0.8802** |
| | 3 cm | **32.83 / 0.9114** | **31.37 / 0.8942** | **23.98 / 0.8377** | **28.98 / 0.8844** | **29.29 / 0.8819** |
| | 5 cm | **31.70 / 0.8935** | **31.03 / 0.8820** | **24.07 / 0.8374** | **28.63 / 0.8686** | **28.86 / 0.8704** |
| MambaIR | 0 cm | 31.37 / 0.8826 | 27.99 / 0.8294 | 21.52 / 0.7350 | 22.97 / 0.7731 | 25.96 / 0.8050 |
| | 1 cm | 30.36 / 0.8743 | 27.86 / 0.8319 | 20.32 / 0.7146 | 23.51 / 0.7778 | 25.51 / 0.7996 |
| | 3 cm | 30.24 / 0.8676 | 27.59 / 0.8226 | 20.82 / 0.7187 | 24.26 / 0.7933 | 25.73 / 0.8006 |
| | 5 cm | 29.11 / 0.8470 | 27.41 / 0.8093 | 20.99 / 0.7127 | 24.24 / 0.7858 | 25.44 / 0.7887 |

Table 3. Comparison of segmentation (mIoU/pAcc) and restoration (PSNR/SSIM) performance with varying proportions of the CLP training set (10%–100%).

| Method | Scenes for Training | | | | |
|---|---|---|---|---|---|
| | 10% | 25% | 50% | 75% | 100% |
| **Semantic Segmentation (mIoU / pAcc)** | | | | | |
| DAFormer | 10.01 / 50.09 | 21.84 / 59.66 | 21.77 / 62.01 | 26.64 / 68.28 | **30.45 / 70.18** |
| SoMA | 14.96 / 57.38 | 24.03 / 59.95 | 32.03 / 69.21 | 37.10 / 71.78 | **43.73 / 73.10** |
| **Image Restoration (PSNR / SSIM)** | | | | | |
| Restormer | 22.93 / 0.7287 | 23.07 / 0.7482 | 23.63 / 0.7598 | 23.63 / 0.7615 | **24.77 / 0.7781** |
| NAFNet | 23.55 / 0.7464 | 23.91 / 0.7547 | 24.26 / 0.7647 | 24.70 / 0.7698 | **25.10 / 0.7834** |

**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 $512 \times 512$ for segmentation and $128 \times 128$ 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 and pixel accuracy, while restoration outputs are evaluated at their original resolution using PSNR and SSIM. All experiments run on a single NVIDIA RTX 4090 GPU with fixed random seeds. Please refer to the supplementary material for each model's detailed hyperparameters.

## 5.2. Semantic Segmentation Results

**Quantitative 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 [49] and Rein [44] achieving improvements of +11.5 and +9.8 mIoU over leading DA baselines. 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 [30] achieves the highest average mIoU (47.4%) across all contamination types, likely due to its large-scale pretraining on diverse data and strong visual prior. In contrast, SAM [20] performs worse despite its scale, reflecting the weaker generalization of its prompt-driven design under distortions and occlusion.

Table 4. Semantic segmentation results (mIoU / pAcc) under combinations of restoration and segmentation models.

| | Restoration | | Segmentation | | mIoU | pAcc |
|---|---|---|---|---|---|---|
| | DiffUIR | NAFNet | DINOv2 | Mask2Former | | |
| (1) | ✗ | ✗ | ✓ | ✗ | 45.4 | 78.6 |
| (2) | ✓ | ✗ | ✓ | ✗ | 45.4 | 74.2 |
| (3) | ✗ | ✓ | ✓ | ✗ | 43.6 | 74.2 |
| (4) | ✗ | ✗ | ✗ | ✓ | 33.4 | 66.8 |
| (5) | ✓ | ✗ | ✗ | ✓ | 33.6 | 68.3 |
| (6) | ✗ | ✓ | ✗ | ✓ | 33.5 | 66.9 |

Table 5. Comparison of restoration performance when models are trained on SIDL vs. CLP (averaged over 0–5cm).

| Method | Training | Condensation | Mud | Water |
|---|---|---|---|---|
| NAFNet | CLP | 23.06 / 0.8452 | 24.62 / 0.8443 | 25.11 / 0.8527 |
| | SIDL | 20.72 / 0.7258 | 17.74 / 0.6429 | 22.70 / 0.7468 |
| | Difference | +2.34 / +0.1194 | +6.88 / +0.2014 | +2.41 / +0.1059 |
| Restormer | CLP | 23.33 / 0.8502 | 24.75 / 0.8698 | 25.28 / 0.8590 |
| | SIDL | 21.19 / 0.7402 | 19.87 / 0.7283 | 22.47 / 0.7481 |
| | Difference | +2.14 / +0.1100 | +4.88 / +0.1415 | +2.81 / +0.1109 |

**Qualitative Results.** Figure 8 provides visual insights that complement the quantitative trends in Table 1. SoMA [49] achieves competitive quantitative performance, but exhibits distinct qualitative characteristics. It generates finer segmentation masks compared to other approaches, maintaining object boundaries even under degraded conditions. DINOv2 achieves superior overall performance, but generates notably coarse predictions under severe occlusion, indicating that even extensive pre-training cannot fully compensate for completely missing visual information. The model appears to default to generalized, low-confidence predictions when faced with extreme uncertainty. Please see the Appendix for additional qualitative examples.

### 5.3. Image Restoration Results

Table 2 and Figure 9 present the quantitative and qualitative comparisons of restoration models on our dataset. DiffUIR [53] achieves the best performance, indicating that diffusion-based models are well suited for handling our challenging contamination types. However, artifacts often remain in severely degraded regions (Figure 9), indicating that a breakthrough will require more advanced models for robust restoration.

### 5.4. Model Complexity Analysis

Figure 10 presents the trade-offs between performance and model complexity for both segmentation and restoration. In segmentation (**left**), parameter-efficient models such as SoMA [49], DINOv2 [30], and Rein [44] achieve higher mIoU per compute compared to larger foundation models like SAM [20], suggesting that efficiency-oriented designs can be more effective under contamination. In contrast, the restoration results (**right**) show no clear relationship between

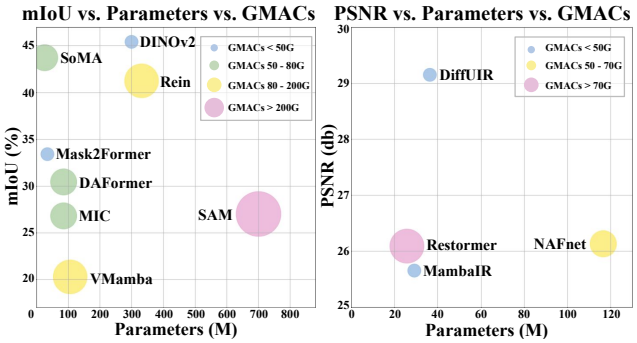


Figure 10. Model complexity *vs.* performance for segmentation (**left**) and restoration (**right**). Segmentation models show a trade-off between accuracy and complexity, while restoration models exhibit no consistent pattern.

PSNR and either model size or computational cost, indicating that future gains may stem from generative approaches (*e.g.,* diffusion-based methods) rather than further scaling of deterministic architectures.

### 5.5. Ablation Studies

**Data Scale.** Table 3 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.** Table 4 shows segmentation performance using combinations of restoration models (DiffUIR [53], NAFNet [6]) and segmentation models (DINOv2 [30], Mask2Former [7]). Each segmentation model was also tested without restoration as a baseline. Interestingly, DINOv2 experienced a performance drop when combined with restoration models, whereas Mask2Former showed performance improvement. This contrast likely stems from the difference between foundation models and supervised models, suggesting an intriguing direction for future investigation.

**SIDL *vs.* CLP** Table 5 quantitatively compares restoration models (NAFNet [6], Restormer [50]) trained on either SIDL or CLP, evaluated on the CLP test set. Although SIDL [9] and CLP are qualitatively similar in terms of image content (see Appendix), models trained on SIDL perform significantly worse on the CLP test set. This performance gap shows that CLP exposes domain shifts not covered by SIDL, making it a more faithful testbed for restoration methods deployed under realistic lens protector contamination.

## 6. Limitation

While CLP provides a controlled and diverse benchmark for segmentation and restoration under real-world contamination, its scope is currently limited to RGB imagery with semantic annotations. The dataset does not include additional modalities such as depth maps or multi-view captures, which could facilitate studies on 3D perception and multi-modal learning under contamination. We plan to expand CLP to include these modalities and increase scale in future work.

## 7. Conclusion

We introduced CLP (Contaminated Lens Protectors), a real-world dataset designed to benchmark robust perception for practical vision systems under realistic sensor contamination. CLP systematically captures multiple types of lens-protector degradations, including mud, water droplets, and condensation, across varied lens-to-protector distances. This captures paired clean and contaminated images, dense semantic segmentation masks, and aligned restoration targets. Our experiments benchmark diverse segmentation approaches, including domain generalization and foundation models, the effect of data scale, and joint restoration–segmentation pipelines. These results highlight strategies for addressing data scarcity and guide the development of reliable perception in real-world contaminated-vision scenarios. We will make the CLP dataset publicly available for the research community.

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

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

[8] 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

[9] 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, 8

[10] 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

[11] 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

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

[13] 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

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

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

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

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

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

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

[20] 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

[21] 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

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

[23] 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

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

[25] 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

[26] 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

[27] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *NeurIPS*, 2024. 5

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

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

[30] 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

[31] 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

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

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

[34] 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

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

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

[37] 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

[38] 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

[39] 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

[40] 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

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

[42] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023. 3

[43] 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

[44] 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, 6, 7, 8

[45] Xin Yang, Xin Zhang, and Xinchao Wang. Erf: A benchmark dataset for robust semantic segmentation under extreme rainfall conditions. In *AAAI*, 2025. 6

[46] 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

[47] 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

[48] 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

[49] 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, 6, 7, 8

[50] 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, 8

[51] 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

[52] 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

[53] 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

[54] 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