2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Train Once, Apply Broadly: Low-Frequency Generative Augmentation for Driver Distraction Recognition under Photometric Shifts

Dichao Liu[1], Longjiao Zhao[2*], Mingkai Gu[5], HaoJiang Chen[4], Ying Ji[3]

Code: https://github.com/Dichao-Liu/ddr-lfga

***Abstract*— Driver distraction recognition (DDR) degrades under deployment-time shifts in camera/ISP pipelines and illumination. We frame this as a single-source domain generalization (SSDG) problem: training on one labeled source domain and testing on unseen devices and lighting. Motivated by this, we propose Low-Frequency Generative Augmentation (LFGA), which separates each image into a fixed high-frequency structure and a re-renderable low-frequency base. Multi-stage, feature-conditioned generators perturb only the photometric low-frequency content and recombine it with the original high-frequency structure to yield "hard-but-correct" views to teach the model photometric invariances. Training imposes decision consistency via cross-entropy and logit matching, and promotes stage-wise separation along class-agnostic factors with a feature-dissimilarity loss. Generators are training-only. On two DDR benchmarks with synthetic cross-photometric shifts and a zero-shot real cross-device video test, LFGA improves cross-domain performance over strong SSDG and DDR baselines while preserving in-domain accuracy.**

I. Introduction

Driver distraction—the diversion of attention from the primary driving task to secondary activities (e.g., phone use, conversations)—poses a major safety risk to road users. In the United States alone, the National Highway Traffic Safety Administration (NHTSA) reported in April 2025 [1] that distraction-related crashes in 2023 caused 3,275 deaths and about 324,819 injuries. Reliable driver distraction recognition (DDR) and timely intervention are therefore critical for road safety. A fast-maturing strand of Intelligent Transportation Systems (ITS), camera-based DDR is moving toward large-scale deployment; however, in commercial deployment, the same model frequently degrades across different capture devices and illumination conditions, as shown in Figure 1.

A principal cause is photometric shift induced by the camera pipeline and lighting: nonlinear, spatially varying, and content-dependent changes arising from automatic white balance and tone curves, exposure/contrast, local shadows/highlights, and sensor/ISP idiosyncrasies [2]. These shifts go well beyond global, linear jitter modeled by common augmentations (e.g., brightness/contrast ColorJitter). Collecting labeled data per device is rarely practical; consequently, robustness to unseen devices and lighting under single-source training is pivotal for DDR deployment.

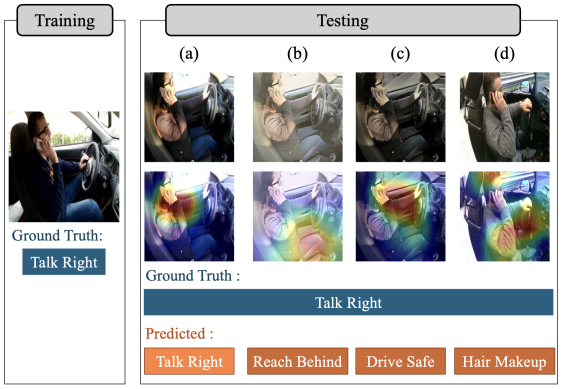

Fig. 1. **Motivation—sensitivity to camera-pipeline appearance shifts.** We evaluate a vanilla MobileNetV2 [3] trained on the AUC Distracted Driver Dataset (AUC-DDD) [4]. Left: one training image. Right: panels (a–c) use the *same* AUC-DDD test image—(a) in-domain (unaltered); (b) photometrically remapped (via neural transfer [5]) to *iPhone 13 in bright sunlight*; (c) photometrically remapped to *Anero Car DVR in low light*; (d) a *real* cross-device frame from the Driver Monitoring Dataset [6] captured with an *Intel RealSense* camera. Top row: inputs; bottom row: Grad-CAM [7], which visualizes the image regions that most influence the target-class decision. The ground truth is the same across panels, while predictions from the same model can change under these shifts—a common barrier to commercial deployment, motivating our work.

This challenge is naturally cast as single-source domain generalization (SSDG): training on a single labeled source while expecting robustness to deployment-time device and illumination shifts, without access to target-domain data.

Along this line of research, data-side augmentation is an active direction for enriching the single-source domain. Broadly, augmentation strategies fall into two streams: (i) adversarially optimized transforms that search for the hardest cases, and (ii) style/statistics mixing that broadens appearance diversity. The former can compromise semantic stability; the latter is often global and content-agnostic, leaving spatially varying, content-coupled factors insufficiently covered. Motivated by this, we propose Low-Frequency Generative Augmentation (LFGA), which constructs "hard-but-correct" views by re-rendering only the low-frequency appearance (e.g., white balance, tone mapping, exposure, contrast, illumination, shadows) while preserving the high-frequency structure (e.g., pose, gesture edges) and keeping the decision unchanged. This encourages the model to learn invariances where deployment-time shifts occur.

LFGA builds on a task property of DDR: driver action semantics reside primarily in high-frequency (HF) structure, whereas device and lighting shifts manifest in low-frequency

[1]The College of Artificial Intelligence, Dalian Maritime University, China
[2]Independent researcher
[3]Wuxi University of Technology, China
[4]Jiangsu University, China
[5]Suzhou University of Science and Technology, China
*Corresponding author: Longjiao Zhao. Email: longjiaozhao@gmail.com.

(LF) appearance. The method first decouples each input into an HF component (kept fixed) and an LF base (mutable). From multiple deep stages of the backbone convolutional neural network (CNN), we extract intermediate representations to condition multiple generators, each producing a photometrically shifted low-frequency (psLF) component. For every stage, we recombine its psLF component with the fixed HF to obtain an LF-augmented view and train on these views. The generators are optimized with two complementary objectives: semantic and decision preservation via cross-entropy and logit consistency, and feature decorrelation at the corresponding stage using a Negative-L1 loss to maximize mid-level representation disparity along class-agnostic factors. This consistent semantics plus maximal representation gap strategy drives variation into device and illumination subspaces while leaving class-defining cues intact. An alternating freeze or unfreeze schedule between the generators and the backbone further reduces classifier sensitivity along these factors and smooths decision boundaries. Generators are used only during training and are removed at inference time.

We evaluated in the single-source setting using the AUC Distracted Driver Dataset (AUC-DDD) [4] and State Farm Dataset (SFD) [8] with standard train/test splits. Our evaluation included the original AUC-DDD/SFD test sets and cross-appearance variants constructed via neural style transfer [5]—commonly used to emulate camera-pipeline shifts (camera/ISP and lighting) in camera-style adaptation [2]: testing images were photometrically remapped into the other dataset's camera/ISP style and into two captured exemplars—*iPhone 13 in bright sunlight* and *Anero Car DVR in low light*—covering representative device and illumination shifts. To examine extension to real video, we performed training-free, zero-shot evaluation on the Driver Monitoring Dataset (DMD) [6], whose capture devices differ and lighting is complex: for randomly selected segments aligned to our class set, we averaged per-frame logits to obtain segment-level predictions. In summary, our contributions are as follows:

- **Problem framing for DDR.** We formalize cross-device/illumination robustness in DDR under the single-source setting and pursue a deployment-oriented SSDG route that seeks "hard-but-correct" training views rather than error-inducing edits.
- **Learning photometric invariance without breaking semantics.** Concretely, we decouple HF/LF, re-render only the LF appearance via multi-stage conditioned generators, and enforce decision consistency while using stage-wise decorrelation to maximize mid-level disparities along class-agnostic factors.
- **Robustness and deployability.** LFGA delivers strong gains on both *photometrically remapped* and *real cross-device* tests over state-of-the-art SSDG methods; generators are training-only and removed at inference, yielding $\sim$3.7 ms per 224×224 frame on Jetson Orin.

II. RELATED STUDIES

A. Driver Distraction Recognition

Driver Distraction Recognition (DDR) has long been a focal component of intelligent driver-assistance systems. Early DDR drew on physiological signals, driving-performance cues, and eye tracking; however, recently, camera-based deep models dominate thanks to low intrusiveness and strong accuracy [9].

Setting photometric shift aside, recent DDR work has pushed accuracy very high. For example, CoViT [10] coupled multi-scale/dilated convolutions and lightweight attention with ViT blocks, while Si-CA MobileNet [11] alternated SimAM and Coordinate Attention in a MobileNet-style backbone; both reported strong performance on SFDD.

Nevertheless, most DDR studies emphasize in-domain accuracy/efficiency and pay limited attention to robustness under a single labeled source when models are moved across devices—a practical deployment pain point. Some works (e.g., JAST [12]) pursue cross-sensor generalization using phone sensors (accelerometers and other time-series) to detect abnormal driving; yet, to the best of our knowledge, cross-domain DDR under photometric shift is sparsely explored in the open literature. In industry, robustness against photometric shift is often sought via cross-device data collection plus large-scale labeling—costly and still insufficient to cover the breadth of real-world domains.

B. Single-Source Domain Generalization

Domain generalization (DG) targets performance on unseen domains without target supervision, typically using multiple labeled sources [13]. For DDR, curating such multi-source data is costly, while domain adaptation assumes access to target-domain data and adapts only to known targets—impractical pre-deployment and unable to anticipate future devices/lighting. Hence the single-source setting is more realistic: SSDG trains on one labeled source and generalize to unseen deployment domains. Prior SSDG work falls into two lines: data-side augmentation (broadening source support) and representation-side regularization (shaping domain-robust features).

Data-side augmentation. Two broad families are common. The first adversarially optimizes transforms to synthesize the "hardest" yet label-preserving samples; for example, AdvST [14] learned parameters of semantic transformations via a mini–max objective that maximized model loss during the transform step and then trained on the resulting challenging views, thus connecting the procedure to distributionally robust optimization. The second broadens appearance diversity by mixing or synthesizing styles/statistics; for instance, IMEC [15] augmented training with language-guided style vectors to target specific domains. Beyond mixing, some methods synthesize unseen-domain samples from a single source to expand coverage; for example, PICF [16] learned invariant causal features via a front-door foreground filter and augmented them with sampled object-irrelevant styles. In this line, our LFGA is a data-side augmentation that—rather than above global style/statistics mixing or adversarial

hardest-case transforms—generates "hard-but-correct" views via structure-preserving re-rendering of LF appearance.

Representation-side regularization. Beyond augmentations, normalization/whitening and related objectives target domain-robust features. For example, ADA [17] learned both standardization and rescaling statistics and, when combined with adversarial domain augmentation, improved cross-domain generalization—underscoring the role of representation-side losses in SSDG.

III. METHOD

A. Problem Statement and Design Principles

We study distracted-driving recognition (DDR) under a "single-domain training $\rightarrow$ cross-device / cross-illumination evaluation" setting. The source domain consists of monitoring images acquired by a single device/environment, while the target domain contains *unseen* devices and strong/weak lighting conditions. In realistic development, the source-domain samples and device coverage are inevitably limited, whereas deployment faces diverse and evolving devices and illumination. Our goal is thus to learn a model that is robust to photometric changes yet sensitive to driving-action cues.

To this end, we propose a self-adaptive domain-diversification strategy based on low-frequency generative augmentation. Conditioned on multi-stage intermediate features, the generators synthesize label-consistent photometric variants by perturbing only the low-frequency appearance, expanding coverage along class-agnostic camera/lighting factors while preserving semantics and decisions. During training, this continually broadens the photometric-domain coverage and forces the model to learn domain-invariant representations on these hard yet label-consistent augmentations. As shown in Figure 2, at deployment the generative branch is removed; inference is a single feed-forward pass using the backbone and classification heads, with no image-synthesis overhead.

Formally, let the input image, normalized to $[-1, 1]$, be $I \in [-1, 1]^{3\times H\times W}$, where H and W denote the spatial height and width (in pixels). We choose a fixed high-pass operator $\mathcal{H}(\cdot)$ instantiated as the standard Sobel gradient magnitude [18], treated as non-trainable with stop-gradient and normalized per image to $[-1, 1]$, to obtain

$$I_{\mathrm{HF}} = \mathcal{H}(I), \qquad I_{\mathrm{LF}} = I - I_{\mathrm{HF}}. \tag{1}$$

Here, I_{HF} denotes the high-frequency (HF) component extracted from I, and I_{LF} denotes the complementary low-frequency (LF) base.

Using Sobel magnitude preserves geometric edges/details while being approximately invariant to low-frequency statistics (e.g., exposure/white balance), so that photometric variation is confined to I_{LF} and subsequent LFGA training can broaden photometric-domain diversity under a semantics-preserving constraint.

B. Low-Frequency Generative Augmentation (LFGA)

We use a MobileNetV2 [3] as our backbone and denote it as $\mathcal{B}$. In a CNN, different layers exhibit varying sensitivities to visual information: shallower layers focus on local textures, mid-level layers on object parts and regional tones, and deeper layers on global semantics and class evidence. To comprehensively explore challenging augmentation directions across multiple semantic levels—thereby constructing a more complete domain perturbation space—we adopt a multi-scale adversarial approach.

Specifically, we extract the feature maps $\{x_1, x_2, x_3\}$ from the last three stages of $\mathcal{B}$. The k-th feature map, x_k, has dimensions of $\mathbb{R}^{C_k\times H_k\times W_k}$, where C_k, H_k, and W_k are the number of channels, height, and width, respectively. For each feature map x_k, we establish a dedicated subsystem comprising a classification head $\mathcal{C}_k$ and a conditional low-frequency generator $\mathcal{G}_k$. In addition to per-stage classification heads, we instantiate a concatenated-feature classifier $\mathcal{C}_{\mathrm{concat}}$ that pools and concatenates features from the last three stages to produce a global prediction and to serve as a stable teacher during training.

Classification head. Each classification head $\mathcal{C}_k$ is designed to map an intermediate representation $x_k \in \mathbb{R}^{C_k\times H_k\times W_k}$ to a logit vector $y_k \in \mathbb{R}^{N_{\mathrm{class}}}$, where N_{class} denotes the number of classes. $\mathcal{C}_k$ first processes x_k with a convolutional block (ConvBlock), consisting of a 1×1 convolution (stride 1, pad 0) for channel expansion followed by a 3×3 convolution (stride 1, pad 1) for spatial feature refinement. The resulting feature map is then reduced to a feature vector via global max pooling (GMP). This vector is passed to a two-layer MLP with an ELU activation in the hidden layer to yield the logits. The entire operation can be summarized as :

$$y_k = \mathrm{MLP}(\mathrm{GMP}(\mathrm{ConvBlock}(x_k))). \tag{2}$$

For the concatenation head, we pool and concatenate the three stage features and obtain

$$y_{\mathrm{concat}} = \mathcal{C}_{\mathrm{concat}}\big(\mathrm{GMP}([x_1, x_2, x_3])\big), \tag{3}$$

which is later used as a stable teacher on the clean image.

Conditional generator. Each generator $\mathcal{G}_k$ is a conditional dual-path network that, given the LF base I_{LF} and conditioned on the stage-k feature x_k, synthesizes a photometrically shifted low-frequency (psLF) component $I^{\mathrm{ps}}_{\mathrm{LF},k}$.

The network consists of two parallel streams:

1) Feature stream: The feature map x_k is processed by a decoder, $\mathcal{D}_k$, composed of a series of M blocks. Each decoder block performs $2\times$ bilinear upsampling, followed by a refinement block consisting of a 3×3 convolution (stride 1, pad 1), instance normalization, and ReLU; we set $M = \lceil \log_2(H/H_k) \rceil = \lceil \log_2(W/W_k) \rceil$ so the decoded feature reaches the LF image resolution $(H \times W)$.
2) *Image stream:* The LF image I_{LF} is fed to a shallow encoder $\mathcal{E}$ composed of a single 5×5 convolution (stride 1, pad 2) followed by InstanceNorm and ReLU.

The outputs of both streams are concatenated along the channel dimension and fed into a final 7×7 convolution (stride 1, pad 3) with a Tanh activation to produce the 3-channel psLF component. The overall process is:

$$I^{\mathrm{ps}}_{\mathrm{LF},k} = \mathrm{Tanh}(\mathrm{Conv}_{7\times7}(\mathrm{concat}[\mathcal{D}_k(x_k), \mathcal{E}(I_{\mathrm{LF}})])). \tag{4}$$

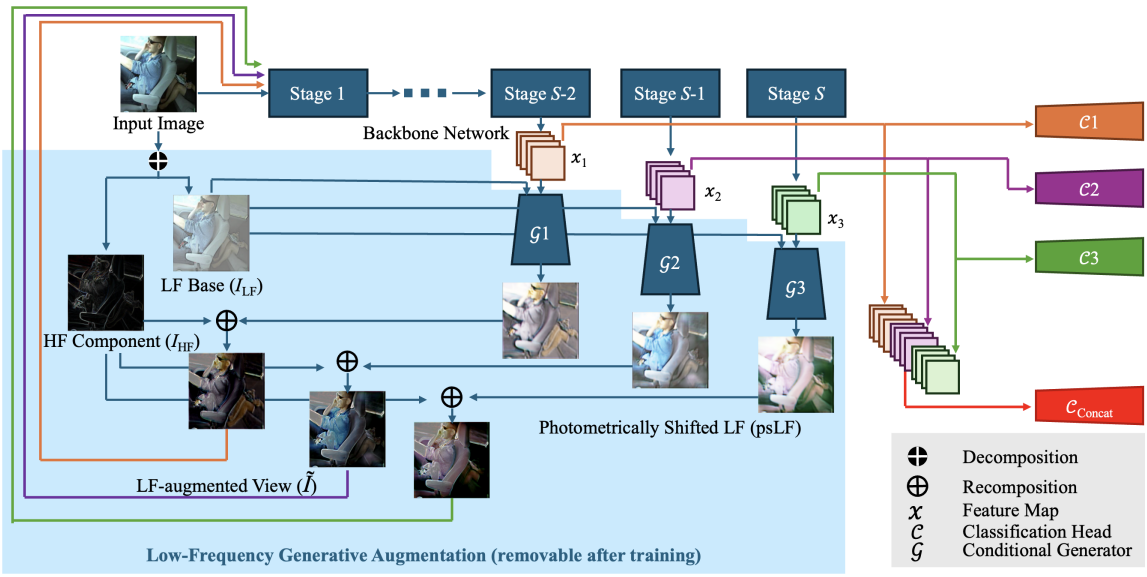

Fig. 2. LFGA pipeline. The input image is decomposed into a high-frequency component I_{HF} and a low-frequency base I_{LF}. At three backbone stages, feature maps x_k condition generators $\mathcal{G}_k$ to synthesize photometrically shifted low-frequency components (psLF), which are recomposed with I_{HF} to form LF-augmented views $\tilde{I}^k$ for training. Color-coded paths indicate training flow: each $\tilde{I}^k$ (orange/purple/green) supervises its matched branch $\mathcal{C}_k$, while the red path denotes the concatenation head $\mathcal{C}_{\mathrm{concat}}$ trained on features from the original image. Conditioning on stage-wise backbone features yields diverse, stage-specific photometric-domain perturbations—from local tone/shadow changes to global color cast—while alternating generator–classifier updates progressively expand coverage of camera/illumination shifts. The blue-shaded region indicates the training-only generative path, removed at inference.

The final augmented image $\tilde{I}_k$ is composed by combining the original HF component I_{HF} with the psLF component:

$$\tilde{I}_k = I_{\mathrm{HF}} + I^{\mathrm{ps}}_{\mathrm{LF},k}. \tag{5}$$

Generator training. The generators $\{\mathcal{G}_k\}$ are trained adversarially with the backbone $\mathcal{B}$ and all classification heads frozen. The objective for each generator $\mathcal{G}_k$ is to synthesize an augmented image $\tilde{I}_k$ that is challenging for the classifier yet semantically consistent. This is achieved through a composite loss function $\mathcal{L}_{\mathcal{G}_k}$ that balances three objectives.

Let $y_j = \mathcal{C}_j(\mathcal{B}(I))$ be the logit vector from the original image and $y'_j = \mathcal{C}_j(\mathcal{B}(\tilde{I}_k))$ be the logit vector from the augmented image, for any stage j. Similarly, let x_k and x'_k be the intermediate feature maps at stage k from the original and augmented images, respectively. The loss components are:

1) *Multi-view Classification Loss ($\mathcal{L}_{\mathrm{cls}}$):* Enforces semantic validity by ensuring the augmented image is correctly classified by other classification heads ($j \neq k$). Let y_{gt} be the ground truth label. This loss is the cross-entropy (CE) loss summed over these other heads:

$$\mathcal{L}_{\mathrm{cls}} = \sum_{j \neq k} \mathcal{L}_{\mathrm{CE}}(y'_j, y_{\mathrm{gt}}). \tag{6}$$

2) *Logit Matching Loss ($\mathcal{L}_{\mathrm{logit}}$):* Imposes a stricter consistency by requiring the logit distributions from the augmented image to match those from the original image, using an L1 distance across other heads ($j \neq k$):

$$\mathcal{L}_{\mathrm{logit}} = \sum_{j \neq k} \|y'_j - y_j\|_1. \tag{7}$$

3) *Feature Dissimilarity Loss ($\mathcal{L}_{\mathrm{feat}}$):* The core adversarial objective, which pushes the representation of the augmented image x'_k away from the original x_k at the target stage k. This is achieved by maximizing the L1 distance between their normalized versions:

$$\mathcal{L}_{\mathrm{feat}} = -\|\mathrm{norm}(x'_k) - \mathrm{norm}(x_k)\|_1. \tag{8}$$

The final loss for training the generator $\mathcal{G}_k$ is the direct sum of these three objective terms:

$$\mathcal{L}_{\mathcal{G}_k} = \mathcal{L}_{\mathrm{cls}} + \mathcal{L}_{\mathrm{logit}} + \mathcal{L}_{\mathrm{feat}}. \tag{9}$$

Classifier training and overall procedure. After updating the generators, we freeze their parameters and train the main recognition model—that is, the backbone $\mathcal{B}$ together with all classification heads. This step is designed to enhance the classifier's robustness by forcing it to learn from the "hard-but-correct" samples synthesized by the generators. The training is divided between the multi-scale branches and the main concatenated branch.

Each branch classifier $\mathcal{C}_k$ is trained on its corresponding augmented image $\tilde{I}_k$. Its loss function, $\mathcal{L}_{\mathcal{C}_k}$, combines two terms. The first is a standard cross-entropy loss with the ground-truth label y_{gt}. The second is a knowledge distillation (KD) loss, $\mathcal{L}_{\mathrm{KD}}$, which regularizes the training. For this, the "teacher" is the stable prediction of the concatenation head $\mathcal{C}_{\mathrm{concat}}$ on the original, clean image I, see Equation (3). The "student" is the branch classifier $\mathcal{C}_k$ observing the augmented image. The loss for branch k is:

$$\mathcal{L}_{\mathcal{C}_k} = \mathcal{L}_{\mathrm{CE}}(y'_k, y_{\mathrm{gt}}) + \mathcal{L}_{\mathrm{KD}}(y'_k, y_{\mathrm{concat}}), \tag{10}$$

where y'_k are the logits from branch k on $\tilde{I}_k$, y_{concat} are the teacher logits from the clean image I. KD constrains

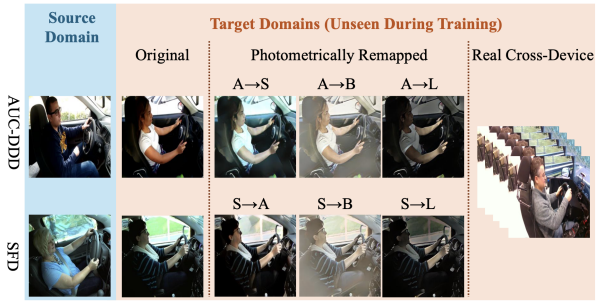

Fig. 3. Left: training source (AUC-DDD or SFD); middle: unseen targets via synthetic photometric remapping (A→S/A→B/A→L, S→A/S→B/S→L; A=AUC-DDD, S=SFD; arrows denote source→target style, B=iPhone 13 in bright sunlight, L=Anero Car DVR in low light); right: real cross-device sample from DMD (Intel RealSense).

the student $\mathcal{C}_k$ on $\tilde{I}_k$ by aligning its logits y'_k to the teacher distribution y_{concat} from $\mathcal{C}_{\text{concat}}$ on the clean input I, thereby preserving semantics under LF augmentation and promoting cross-branch consistency.

Concurrently, the concatenation head $\mathcal{C}_{\text{concat}}$ is trained exclusively on the original image I using only the standard cross-entropy loss, ensuring it remains a reliable anchor to the source data distribution.

The overall training follows a two-stage alternating optimization procedure for each batch of data:

1) *Generator Update:* First, freeze the backbone $\mathcal{B}$ and all classifier heads, and update the parameters of all generators $\{\mathcal{G}_k\}$ by minimizing their composite loss $\mathcal{L}_{\mathcal{G}_k}$ in Equation (9).
2) *Classifier Update:* Next, freeze the generators $\{\mathcal{G}_k\}$ and update the backbone $\mathcal{B}$ and all classifier heads. This is done by minimizing the sum of the per-branch losses $\{\mathcal{L}_{\mathcal{C}_k}\}$ in Equation (10), together with the classification loss of $\mathcal{C}_{\text{concat}}$, defined as $\mathcal{L}_{\text{CE}}(y_{\text{concat}}, y_{\text{gt}})$.

This iterative, adversarial process enables the generators and the classifier to progressively improve, leading to a final model with enhanced generalization capabilities. At inference, we average the logits from the three stage heads and the concatenation head to obtain the final prediction.

IV. Evaluation

A. Datasets and Evaluation Protocol

Source datasets. We adopted two image-based DDR benchmarks as the single labeled source: the AUC Distracted Driver (AUC-DDD) [4] and the State Farm Distracted Driver (SFD) [8]. Both are RGB image datasets with a 10-class label space: safe driving, texting right, talking on the phone right, texting left, talking on the phone left, adjusting radio, drinking, reaching behind, hair and makeup, and talking to passenger. AUC-DDD contains 17,308 images with an official split (12,977 train / 4,331 test); we trained on the official training portion and held out a validation subset. The dataset was recorded with an ASUS ZenPhone. SFD provides 22,424 labeled training images and 79,728 unlabeled test images; following common practice [9], we randomly split the labeled set into 70% training / 30% testing and report results on the held-out test set and on the external cross-appearance tests. Although SFD does not disclose its capture device, as shown in Figure 3, a clearly different camera/ISP rendering relative to AUC-DDD (e.g., white balance and tone mapping) indicates a device-induced photometric shift.

External evaluation under synthetic photometric shifts. To assess robustness to camera/ISP and illumination shifts under the single-source constraint, we applied photometric remapping via neural style transfer [5], producing synthetic cross-appearance counterparts as a commonly used proxy for camera-pipeline (camera/ISP) and lighting changes [2]. As illustrated in Figure 3, when AUC-DDD was the source, its testing images were synthetically remapped to (i) the SFD camera/ISP rendering ($A \to S$) and to two exemplar-guided styles; (ii) iPhone 13 in bright sunlight ($A \to B$) and (iii) Anero Car DVR in low light ($A \to L$). When SFD was the source, we analogously synthesized remapped counterparts to the AUC-DDD rendering ($S \to A$) and to the same two exemplar-guided styles ($S \to B$, $S \to L$). We manually reviewed all generated images to ensure semantics were preserved and no artifacts or label noise were introduced. In-domain results were reported on the unaltered testing splits.

Zero-shot evaluation on real cross-device video. To probe transfer to real videos with unseen devices and complex lighting, we performed zero-shot evaluation on the Driver Monitoring Dataset (DMD) [6]—with no training, fine-tuning, or calibration on DMD—making this a particularly stringent out-of-domain test. DMD is a driver-monitoring video corpus (∼41 hours) for attention/arousal estimation, covering distraction, drowsiness, gaze, and related behaviors. For our focus on distraction, we restricted evaluation to clips aligned with our 10-class label space: because DMD includes additional categories and is highly imbalanced, we retained only the 10 overlapping classes and randomly sampled 40 clips per class (total 400 clips). For this experiment we used AUC-DDD as the single labeled source (it offers greater intra-class diversity than SFD), and all baselines followed the same zero-shot protocol. For each selected clip, we ran frame-wise inference and averaged per-frame logits to obtain the clip-level prediction.

Training details. We trained for 100 epochs with batch size 16 on an NVIDIA GeForce RTX 4080 using alternating optimization—updating the low-frequency generators $\mathcal{G}_1$–$\mathcal{G}_3$ with Adam (initial learning rate 2×10^{-4}, betas $0.5, 0.999$) and the backbone plus classification heads with Stochastic Gradient Descent (initial learning rate 2×10^{-3}, momentum 0.9, weight decay 5×10^{-4}) under a cosine-annealing schedule; training images were resized to 256×256 then randomly cropped to 224×224 with horizontal flip, and evaluation used resize and center crop to 224×224; the number of conditioned backbone stages K was set to 3; we reported Accuracy (Acc), F1-score (F1), Precision (PRE), and Recall (REC); for embedded deployment, the generators were removed and we additionally measured deployment cost (latency and power) on NVIDIA Jetson AGX Orin (32GB). Each experiment was run five times with different random seeds, and we report the mean $\pm$ standard deviation.

TABLE I

COMPARISON WITH STATE-OF-THE-ART METHODS UNDER IN-DOMAIN EVALUATION WITHOUT SYNTHETIC SHIFTS ON AUC-DDD AND SFD.

Method	AUC-DDD				SFD			
	Acc (%)	F1 (%)	PRE (%)	REC (%)	Acc (%)	F1 (%)	PRE (%)	REC (%)
• Baseline (MobileNetV2)								
–	95.3±2.3	95.6±2.3	95.7±2.1	95.1±1.8	99.7±0.1	99.7±0.1	99.7±0.1	99.7±0.1
• Baseline + ColorJitter & Random brightness								
–	95.2±2.2	95.2±2.4	95.5±2.1	95.0±1.8	99.7±0.1	99.7±0.1	99.7±0.1	99.7±0.1
• Baseline + Single-Source Domain Generalization (SSDG)								
ADA	94.2±1.2	94.4±0.9	94.6±1.4	94.2±0.9	99.6±0.2	99.5±0.2	99.5±0.2	99.5±0.2
AdvST	95.1±1.2	95.3±1.3	95.8±1.2	94.8±1.0	**99.8±0.0**	**99.8±0.0**	**99.8±0.0**	**99.8±0.0**
CFA	95.6±1.3	95.7±1.2	95.9±1.2	95.5±1.1	99.7±0.1	99.7±0.1	99.7±0.1	99.7±0.1
DFQ	95.1±1.3	95.2±1.2	95.6±1.2	95.0±1.1	99.7±0.1	99.7±0.1	99.7±0.1	99.7±0.1
PICF	94.2±1.3	94.4±1.2	94.8±1.2	94.0±1.1	99.5±0.2	99.5±0.2	99.5±0.2	99.5±0.2
$\hat{I}S{\rightarrow}I^{\cdot 75}S$	94.5±0.2	94.7±0.1	94.7±0.2	94.7±0.1	99.6±0.2	99.5±0.2	99.6±0.2	99.5±0.2
IMEC	66.8±8.1	67.7±7.4	73.0±4.3	68.8±6.5	78.1±10.4	72.3±11.0	80.9±5.4	72.7±10.6
LFGA (Ours)	**95.8±0.4**	**96.0±0.4**	**96.3±0.6**	**95.7±0.3**	99.8±0.0	99.7±0.0	99.8±0.0	99.7±0.0
• DDR-specific networks (non-SSDG)								
CoViT	95.3±2.6	95.6±2.7	95.8±2.5	95.4±2.4	99.7±0.1	99.7±0.1	99.7±0.1	99.7±0.1
Si-CA	95.3±2.9	95.6±2.8	95.8±2.7	**95.7±2.6**	99.5±0.1	99.5±0.1	99.5±0.1	99.5±0.1

B. Experimental Results

To assess the effectiveness of our approach, we benchmarked against two families of methods. The first comprises state-of-the-art (SOTA) SSDG methods—the category our method also belongs to. To isolate the contribution of augmentation/regularization from backbone capacity, all SSDG methods, including our LFGA, were instantiated on the same MobileNetV2 backbone, which prior work [19] has shown effective for DDR. Concretely, we compared LFGA against ADA [17], AdvST [14], CFA [20], DFQ [21], PICF [16], $\hat{I}S{\rightarrow}I^{\cdot 75}S$ [22], and IMEC [15] under identical settings.

Because our main task is DDR, we additionally included two recent task-specialized DDR models that focus on in-domain accuracy rather than cross-device photometric robustness: CoViT [10] and Si-CA [11].

All methods used identical data splits, hyperparameters, and protocol, ensuring fair and reproducible comparison.

In domain results on original test sets without photometric shifts. Table I shows that, without photometric shifts, performance on both image datasets (AUC-DDD and SFD) is near-saturated. On SFD, many methods reach a ceiling of $\geq 99.5\%$; on AUC-DDD, LFGA is best or tied-best across all metrics. Notably, when added to the MobileNetV2 baseline, several SSDG methods actually reduce performance—e.g., IMEC, PICF, and $\hat{I}S{\rightarrow}I^{\cdot 75}S$—suggesting that these methods, though strong on other single-domain shifts, may not transfer well to the photometric-shifted DDR setting. By contrast, LFGA does not cause in-domain degradation and slightly lifts this ceiling, while matching or modestly exceeding DDR-specific networks (CoViT, Si-CA). Overall, existing networks already yield very high in-domain accuracy; LFGA preserves or marginally improves it; the robustness differences become evident in the more challenging cross-photometric experiments presented below.

Robustness under synthetic photometric shifts. Tables II and III show that under single-source training and photometric-remapping evaluation emulating camera/ISP and illumination shifts, LFGA consistently leads with either training dataset (AUC-DDD or SFD). With AUC-DDD as source (A→S/B/L), LFGA achieves the best performance in all metrics, clearly surpassing ColorJitter and multiple SSDG methods, and it retains a clear advantage even in the most challenging bright-sunlight case (A→B). With SFD as source (S→A/B/L), LFGA also yields stable gains over other methods. DDR-specific networks such as CoViT and Si-CA collapse under cross-photometric shifts.

Zero-shot cross-device evaluation on real video. Table IV reports zero-shot, segment-level evaluation on DMD, where frame-wise inference is aggregated by segment-level logit averaging without any DMD-specific adaptation. This setting is highly challenging, involving cross-device variation, complex illumination, and long video segments. LFGA achieves the best overall performance, with accuracy and F1 around 68–69%, precision in the mid-70% range, and recall closely tracking accuracy. In contrast, most generic SSDG baselines remain in the mid-40% to mid-50% range, while DDR-specific architectures, namely CoViT and Si-CA, degrade further in this difficult scenario. These results highlight that LFGA's principle of "LF-only re-rendering with HF structure and decision consistency" yields stable and significant single-source gains across both synthetic appearance remapping and real-world cross-device video.

Ablation study. Table V reports ablations of LFGA in the single-source setting with AUC-DDD as source and three cross-appearance target domains (A→S, A→B, A→L). We examine three aspects: the number of conditioned generator stages K, the contribution of the generators themselves, and the role of the KD constraint described in Section III-B.

First, varying K shows that a single generator branch yields only marginal improvement over the baseline. Increasing to two or three branches brings consistent gains across all three target domains, with $K = 3$ achieving the best balance between accuracy and stability. Adding more branches ($K = 4, 5$) provides no further benefit and sometimes slightly reduces performance, indicating that three stages already cover the major LF appearance variations.

TABLE II

ROBUSTNESS COMPARISON WITH STATE-OF-THE-ART METHODS UNDER SYNTHETIC PHOTOMETRIC SHIFTS ON AUC-DDD.

Method	A → S				A → B				A → L			
	Acc (%)	F1 (%)	PRE (%)	REC (%)	Acc (%)	F1 (%)	PRE (%)	REC (%)	Acc (%)	F1 (%)	PRE (%)	REC (%)
• Baseline (MobileNetV2)												
–	81.5±2.3	81.7±2.3	84.6±2.1	81.2±1.8	68.8±6.8	68.9±6.9	74.9±3.6	69.4±6.8	88.1±5.6	88.4±5.5	92.0±2.8	87.9±5.9
• Baseline + ColorJitter & Random brightness												
–	86.5±2.2	87.3±2.4	89.0±2.1	86.0±1.8	78.1±6.8	78.0±6.9	80.0±3.6	77.4±6.8	82.9±5.6	83.0±5.7	83.9±2.8	82.8±5.9
• Baseline + Single-Source Domain Generalization (SSDG)												
ADA	80.6±1.2	80.3±0.9	81.2±1.4	80.7±0.9	74.9±3.3	75.3±3.7	76.8±1.5	75.8±3.5	82.1±3.1	81.9±2.6	82.9±1.6	81.5±2.8
AdvST	86.0±1.2	86.7±1.3	88.9±1.2	86.0±1.0	81.7±3.6	82.5±3.7	85.4±2.0	80.8±3.5	80.3±2.9	79.9±2.6	83.9±1.6	78.8±3.1
CFA	84.6±1.3	85.0±1.2	85.5±1.2	85.1±1.1	66.3±3.6	66.6±3.7	76.8±2.0	65.8±3.5	76.3±2.9	76.7±2.7	80.1±1.6	75.5±3.0
DFQ	82.0±1.3	81.4±1.2	84.0±1.2	80.7±1.1	75.3±3.6	74.8±3.7	78.8±2.0	74.2±3.5	76.3±2.9	76.4±2.8	80.8±1.6	74.9±3.0
PICF	82.6±1.3	82.9±1.2	84.5±1.2	82.1±1.1	61.3±3.6	60.3±3.7	70.7±2.0	58.6±3.5	66.1±2.9	66.3±2.8	74.9±1.6	63.6±3.0
$\hat{I}S \rightarrow I^{\cdot 75}S$	68.5±1.5	68.5±1.4	74.4±0.2	67.6±1.8	52.7±1.4	51.5±1.7	59.5±4.9	51.7±2.4	54.0±2.0	52.2±2.4	60.6±1.6	53.0±3.5
IMEC	69.8±1.2	69.8±0.6	75.9±2.9	68.6±0.9	59.2±3.4	58.7±2.4	65.9±1.4	59.1±3.2	55.3±3.3	54.5±4.2	61.8±2.9	55.0±3.8
LFGA (Ours)	**93.4±0.4**	**93.4±0.4**	**93.8±0.6**	**93.3±0.3**	**86.5±2.8**	**86.9±3.0**	**89.9±1.5**	**86.2±3.3**	**91.5±1.1**	**92.1±0.9**	**93.2±0.4**	**91.3±1.3**
• DDR-specific networks (non-SSDG)												
CoViT	47.4±2.6	45.2±2.7	60.2±2.5	46.8±2.4	16.6±3.8	12.8±3.9	60.0±2.1	15.5±3.6	40.7±5.4	35.0±5.6	57.1±3.0	37.6±5.7
Si-CA	62.6±2.9	61.0±2.8	63.0±2.7	61.4±2.6	37.3±4.1	35.1±4.0	45.1±2.4	36.9±3.9	65.3±5.4	65.0±5.6	70.0±3.1	63.9±5.7

TABLE III

ROBUSTNESS COMPARISON WITH STATE-OF-THE-ART METHODS UNDER SYNTHETIC PHOTOMETRIC SHIFTS ON SFD.

Method	S → A				S → B				S → L			
	Acc (%)	F1 (%)	PRE (%)	REC (%)	Acc (%)	F1 (%)	PRE (%)	REC (%)	Acc (%)	F1 (%)	PRE (%)	REC (%)
• Baseline (MobileNetV2)												
–	76.8±6.6	77.8±6.0	88.9±2.2	74.9±6.7	71.1±8.8	71.6±8.4	86.4±2.8	71.5±8.7	81.0±7.2	81.5±7.0	87.5±3.0	81.4±7.0
• Baseline + ColorJitter & Random brightness												
–	91.2±1.6	90.9±1.5	91.8±1.5	91.1±1.4	79.2±3.9	80.5±4.0	89.3±2.3	78.9±3.8	75.4±3.2	76.0±3.1	81.1±1.9	75.2±3.3
• Baseline + Single-Source Domain Generalization (SSDG)												
ADA	94.5±1.3	94.3±1.2	94.8±1.3	94.6±1.2	87.3±3.2	87.7±3.6	89.1±2.0	88.1±3.3	94.6±2.9	94.2±2.6	94.6±1.5	94.7±3.0
AdvST	83.7±1.2	84.6±1.3	89.7±1.3	84.3±1.1	77.1±3.5	76.9±3.6	84.7±2.0	77.2±3.5	77.9±2.9	77.8±2.8	84.9±1.5	78.0±3.0
CFA	95.6±1.3	95.6±1.2	96.1±1.3	95.6±1.1	40.6±3.6	40.6±3.7	86.2±2.0	41.0±3.5	92.8±2.9	93.0±2.6	94.1±1.5	93.0±3.0
DFQ	87.3±1.3	88.4±1.2	92.2±1.2	87.6±1.1	75.3±3.6	74.3±3.7	82.6±2.0	75.8±3.5	85.7±2.9	86.4±2.8	90.2±1.6	85.8±3.0
PICF	91.7±1.3	91.4±1.2	92.3±1.2	91.6±1.1	49.6±3.6	46.0±3.7	72.2±2.0	50.3±3.5	75.9±2.9	76.5±2.8	81.6±1.6	75.7±3.0
$\hat{I}S \rightarrow I^{\cdot 75}S$	91.2±3.9	91.2±3.7	92.5±2.2	91.3±3.8	69.6±8.1	68.9±9.5	82.3±2.9	69.9±8.4	84.5±6.8	84.3±7.3	87.8±3.3	84.0±7.1
IMEC	83.9±6.3	84.6±5.8	88.5±2.7	84.1±6.3	57.9±7.9	58.2±9.0	80.6±3.6	58.2±7.2	78.9±4.0	79.4±3.9	85.1±2.1	78.7±3.9
LFGA (Ours)	**99.2±0.2**	**99.2±0.2**	**99.2±0.3**	**99.2±0.2**	**93.1±2.5**	**93.1±2.7**	**94.9±1.5**	**93.3±2.4**	**98.9±0.5**	**98.9±0.5**	**98.9±0.4**	**98.8±0.5**
• DDR-specific networks (non-SSDG)												
CoViT	52.2±2.8	56.0±2.7	81.4±2.6	53.3±2.5	18.5±3.9	11.6±3.8	40.6±2.2	18.1±3.7	19.5±5.3	18.3±5.5	62.1±3.0	21.4±5.6
Si-CA	80.6±2.7	81.1±2.6	84.2±2.5	81.3±2.4	18.7±3.9	13.9±4.0	50.2±2.3	19.2±3.8	67.3±5.6	68.2±5.5	77.9±3.0	67.7±5.7

TABLE IV

COMPARISON WITH STATE-OF-THE-ART METHODS UNDER ZERO-SHOT CROSS-DEVICE EVALUATION ON REAL VIDEOS FROM DMD.

Method	Acc (%)	F1 (%)	PRE (%)	REC (%)
• Baseline (MobileNetV2)				
–	58.44±1.37	58.75±1.36	66.16±1.03	58.44±1.37
• Baseline + ColorJitter & Random brightness				
–	57.78±1.37	55.41±1.43	64.88±0.77	57.78±1.37
• Baseline + Single-Source Domain Generalization (SSDG)				
ADA	45.67±0.47	44.05±0.45	55.65±0.80	45.67±0.47
AdvST	58.56±2.64	57.38±2.75	61.43±3.29	58.56±2.64
CFA	59.00±2.60	59.67±2.60	71.51±1.51	59.00±2.60
DFQ	59.44±1.66	57.84±2.06	64.23±2.06	59.44±1.66
PICF	50.11±1.64	49.26±1.75	60.01±2.25	50.11±1.64
$\hat{I}S \rightarrow I^{\cdot 75}S$	55.56±2.20	56.72±2.25	63.86±2.72	55.56±2.20
IMEC	43.33±2.13	44.19±2.23	55.42±1.09	43.33±2.13
LFGA (Ours)	**68.00±0.82**	**68.10±0.74**	**74.98±0.30**	**68.00±0.82**
• DDR-specific networks (non-SSDG)				
CoViT	32.89±1.91	33.89±3.09	51.98±4.95	32.89±1.91
Si-CA	47.78±2.08	46.38±1.93	51.67±1.54	47.78±2.08

Second, removing the generators while keeping the multi-head structure ($K = 3$ w/o generators) leads to a clear drop in all metrics, confirming that the photometric perturbations synthesized in the LF space are the primary source of robustness. This demonstrates that LFGA's design of re-rendering only the LF component while preserving the high-frequency semantics is essential for cross-photometric generalization.

Finally, turning off KD constraints ($K = 3$ w/o KD) produces an appreciable degradation compared to the default, and variance increases across domains. This indicates that KD serves as a stabilizer by anchoring branch predictions to the concatenation head on clean images, improving consistency without being the main driver of performance.

Overall, the ablation results support that LFGA achieves robustness by introducing controlled LF diversity through multi-stage generators, while KD alignment provides additional training stability.

Edge deployment on NVIDIA Jetson. We also deployed our model on an NVIDIA Jetson AGX Orin developer kit using TensorRT inference engine. With batch size 1 and 224×224 input (FP16), the per frame latency (p50 / p99) was 3.70 / 3.99 ms, the sustained frame rate was 267 FPS and the average power consumption was 5.8 W, demonstrating strong

TABLE V
RESULTS OF ABLATION EXPERIMENTS.

	A → S				A → B				A → L			
Method	Acc (%)	F1 (%)	PRE (%)	REC (%)	Acc (%)	F1 (%)	PRE (%)	REC (%)	Acc (%)	F1 (%)	PRE (%)	REC (%)
Baseline	81.5±2.3	81.7±2.3	84.6±2.1	81.2±1.8	68.8±6.8	68.9±6.9	74.9±3.6	69.4±6.8	88.1±5.6	88.4±5.5	92.0±2.8	87.9±5.9
Number of backbone stages K used for LFGA conditioning												
K=1	81.8±2.3	82.1±2.4	84.9±2.2	81.6±1.9	69.2±6.9	69.3±6.8	75.3±3.7	69.8±6.7	88.5±5.7	88.7±5.6	92.3±2.9	88.3±5.8
K=2	92.4±0.5	92.4±0.5	92.8±0.6	92.3±0.4	85.5±2.9	85.9±3.1	88.9±1.6	85.2±3.4	90.5±1.2	91.1±1.0	92.2±0.5	90.3±1.4
K=3 (default)	**93.4±0.4**	**93.4±0.4**	**93.8±0.6**	**93.3±0.3**	**86.5±2.8**	**86.9±3.0**	**89.9±1.5**	**86.2±3.3**	**91.5±1.1**	**92.1±0.9**	**93.2±0.4**	**91.3±1.3**
K=4	93.1±0.5	93.1±0.5	93.5±0.6	93.0±0.4	86.2±2.9	86.6±3.1	89.6±1.6	85.9±3.4	91.2±1.2	91.8±1.0	92.9±0.5	91.0±1.4
K=5	93.2±0.5	93.2±0.5	93.6±0.6	93.1±0.4	86.3±2.9	86.7±3.1	89.7±1.6	86.0±3.4	91.3±1.2	91.9±1.0	93.0±0.5	91.1±1.4
Key controls (default K=3)												
K=3 w/o KD	92.6±1.3	92.7±1.2	93.4±1.2	92.3±1.0	84.7±3.5	85.5±3.6	88.9±2.0	85.1±3.5	90.4±3.0	91.2±2.9	92.7±1.6	90.2±3.1
K=3 w/o generators	89.0±2.8	89.9±2.7	92.1±2.9	88.4±2.8	82.8±6.7	84.2±6.8	88.6±3.7	83.4±6.9	84.6±5.7	86.0±5.6	88.7±3.0	84.3±5.8

edge deployment efficiency and real-time performance.

V. CONCLUSION

LFGA addresses a core deployment obstacle for camera-based DDR under single-source training: sensitivity to device and illumination shifts. It decouples structure from appearance and re-renders only the low-frequency component via multi-stage, feature-conditioned generators, yielding "hard-but-correct" views along class-agnostic photometric factors. A consistency-plus-diversity objective combining cross-entropy, logit matching, and a stage-wise feature dissimilarity loss pushes invariance; the generators are used only during training and can be removed at inference. Across robustness under synthetic photometric shifts and zero-shot cross-device evaluation on real video, LFGA delivers substantial and consistent gains while preserving in-domain accuracy.

AI Tool Usage Statement: We used ChatGPT solely for English language editing and grammar refinement. The tool was not used to generate scientific content, or figures.

REFERENCES

[1] National Center for Statistics and Analysis, "Distracted Driving in 2023," National Highway Traffic Safety Administration, Washington, DC, Research Note DOT HS 813 703, apr 2025, summary of Statistical Findings. [Online]. Available: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813703

[2] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5157–5166.

[3] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[4] Y. Abouelnaga, H. M. Eraqi, and M. N. Moustafa, "Real-time distracted driver posture classification," in *Machine Learning for Intelligent Transportation Systems Workshop, 32nd Conference on Neural Information Processing Systems*, Montréal, Canada, 2018.

[5] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*, 2016, pp. 694–711.

[6] J. D. Ortega, N. Kose, P. Cañas, M.-A. Chao, A. Unnervik, M. Nieto, O. Otaegui, and L. Salgado, "Dmd: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis," in *European Conference on Computer Vision*, 2020, pp. 387–405.

[7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[8] A. Montoya, D. Holman, S. D. Science, T. Smith, and W. Kan, "State Farm Distracted Driver Detection," Kaggle competition. Available at: https://www.kaggle.com/competitions/state-farm-distracted-driver-detection, Kaggle, 2016.

[9] D. Tan, W. Tian, C. Wang, L. Chen, and L. Xiong, "Driver distraction behavior recognition for autonomous driving: Approaches, datasets and challenges," *IEEE Transactions on Intelligent Vehicles*, 2024.

[10] Z. Li, X. Zhao, F. Wu, D. Chen, and C. Wang, "A lightweight and efficient distracted driver detection model fusing convolutional neural network and vision transformer," *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[11] M. Lv, Y. Liu, Z. Zha, X. Zheng, H. Wang, Y. Wen, and Z. Guo, "Si-ca mobilenet: A lightweight and efficient convolutional neural network for distracted driver detection," *Neurocomputing*, p. 131281, 2025.

[12] X. Chen, R. Qu, and F. Zhao, "Deep unsupervised transfer adversarial network for abnormal driving behavior recognition based on smartphone sensors," *IEEE Sensors Journal*, 2024.

[13] B. Li, Z. Xu, J. Li, X. Liu, J. Fang, X. Li, and H. Yu, "V2x-dg: Domain generalization for vehicle-to-everything cooperative perception," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 8781–8787.

[14] G. Zheng, M. Huai, and A. Zhang, "Advst: Revisiting data augmentations for single domain generalization," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 19, 2024, pp. 21 832–21 840.

[15] C. Jiang, J. Zhao, J. Deng, Z. Li, and H. Zhang, "Imbuing, enrichment and calibration: Leveraging language for unseen domain extension," *International Journal of Computer Vision*, pp. 1–27, 2025.

[16] Y. Wang, M. Yang, A. Wu, and C. Deng, "Progressive invariant causal feature learning for single domain generalization," *IEEE Transactions on Image Processing*, 2025.

[17] X. Fan, Q. Wang, J. Ke, F. Yang, B. Gong, and M. Zhou, "Adversarially adaptive normalization for single domain generalization," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2021, pp. 8208–8217.

[18] Z. Jin-Yu, C. Yan, and H. Xian-Xiang, "Edge detection of images based on improved sobel operator and genetic algorithms," in *2009 International Conference on Image Analysis and Signal Processing*, 2009, pp. 31–35.

[19] D. Srivastava, P. Shah, and S. Shaikh, "Driver activity monitoring using mobilenets," in *Emerging Research in Computing, Information, Communication and Applications: ERCICA 2020, Volume 1*. Springer, 2021, pp. 49–58.

[20] X. Guo, C. Liu, X. Qian, Z. Wang, X. Feng, and Y. Xue, "Single-domain generalized object detection with frequency whitening and contrastive learning," *IEEE Transactions on Multimedia*, 2025.

[21] Q. Bi, J. Yi, H. Zheng, W. Ji, Y. Huang, Y. Li, and Y. Zheng, "Learning generalized medical image representation by decoupled feature queries," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[22] N. Efthymiadis, G. Tolias, and O. Chum, "Crafting distribution shifts for validation and training in single source domain generalization," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025, pp. 1883–1892.