A APPENDIX

A.1 OVERALL PIPELINE FIGURE

We show the overall pipeline of our method in Fig. 7. We first generate perturbed datasets offline with sensitivity analysis. Then during the online process, we conduct adversarial training for each perturbation and pick one level with the worst validation performance, then combine all the selected datasets from all perturbations together with the base dataset, and use them to train the backbone network iteratively.

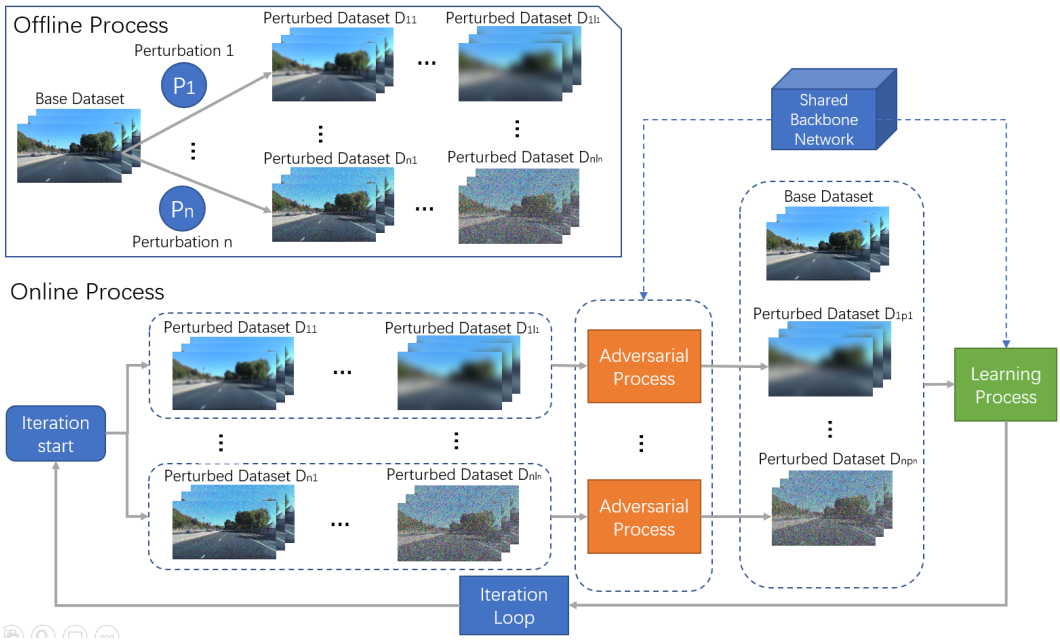

Figure 7: Overall pipeline of our method.

A.2 DATASET SAMPLES

We show different kinds of perturbations in our benchmarks in Fig.8. Specifically, our benchmarks include 9 basic types of perturbations, including Gaussian blur, Gaussian noise, radial distortion, and RGB and HSV channels. Another type of datasets include multiple perturbations, where we create multiple random combinations of the basic perturbations. We also include 7 types of previously unseen perturbations (during training) from ImageNet-C, which are snow, fog, frost, motion blur, zoom blur, pixelate, and jpeg compression. For each type of perturbation, we generate 5 or 10 levels of varying intensity based on sensitivity analysis in the FID-MA space.

A.3 PERTURBED DATASETS

The final representative datasets from the sensitivity analysis and used for improving the generalization of the learning task are introduced in the following.

- R: the base dataset (Chen, 2018);
- $B1, B2, B3, B4, B5$: add Gaussian blur to R with standard deviation $\sigma = 1.4$, $\sigma = 2.9$, $\sigma = 5.9$, $\sigma = 10.4$, $\sigma = 16.4$, which are equivalent to using the kernel (7, 7), (17, 17), (37, 37), (67, 67), (107, 107), respectively;
- $N1, N2, N3, N4, N5$: add Gaussian noise to R with $(\mu = 0, \sigma = 20)$, $(\mu = 0, \sigma = 50)$, $(\mu = 0, \sigma = 100)$, $(\mu = 0, \sigma = 150)$, $(\mu = 0, \sigma = 200)$, respectively;

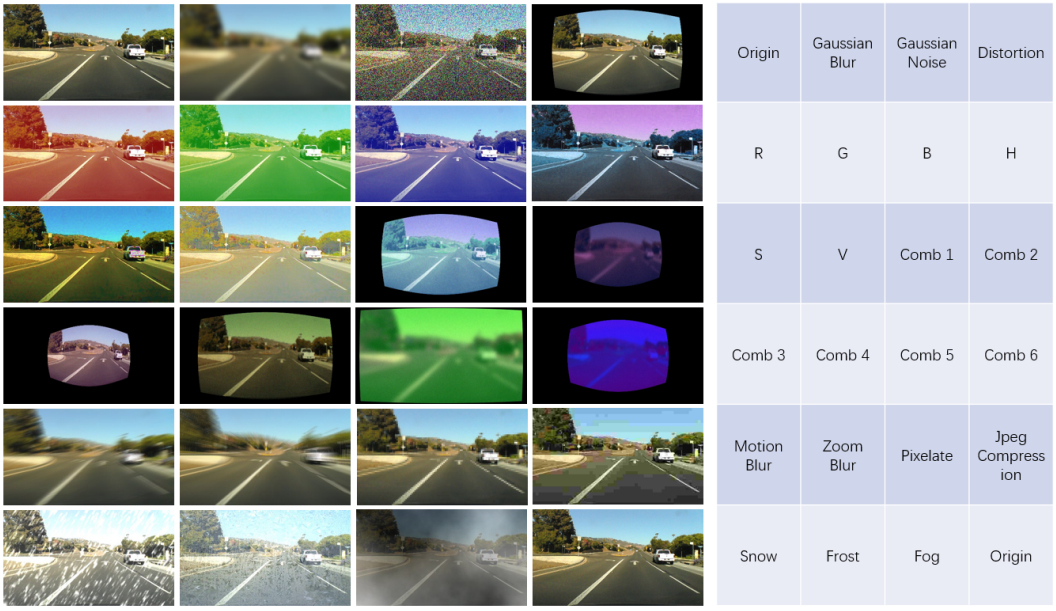

Figure 8: Sample images of our benchmark. We show our benchmark has 22 different types of perturbations. Also, we have 10 levels for R, G, B, H, S, V (5 levels in darker and 5 levels in lighter shades), and 5 levels for each of the other types of perturbations.

- $D1, D2, D3, D4, D5$: distort R with the radial distortion $(k_1 = 1, k_2 = 1)$, $(k_1 = 10, k_2 = 10)$, $(k_1 = 50, k_2 = 50)$, $(k_1 = 200, k_2 = 200)$, $(k_1 = 500, k_2 = 500)$, respectively. k_1 and k_2 are radial distortion parameters, the focal length is 1000, and the principle point is the center of the image.
- $RD1/RL1, RD2/RL2, RD3/RL3, RD4/RL4, RD5/RL5$: modify the red channel of R to darker (D) / lighter (L) values with $\alpha = 0.02$, $\alpha = 0.2$, $\alpha = 0.5$, $\alpha = 0.65$, $\alpha = 1$.
- $GD1/GL1, GD2/GL2, GD3/GL3, GD4/GL4, GD5/GL5$: modify the green channel of R to darker (D) / lighter (L) values with $\alpha = 0.02$, $\alpha = 0.2$, $\alpha = 0.5$, $\alpha = 0.65$, $\alpha = 1$.
- For B, H, S, V channels, we use similar naming conventions for notation as for the red and green channels.
- $Comb1$: $R_\alpha = -0.1180$, $G_\alpha = 0.4343$, $B_\alpha = 0.1445$, $H_\alpha = 0.3040$, $S_\alpha = -0.2600$, $V_\alpha = 0.1816$, $Blur_\sigma = 3$, $Noise_\sigma = 10$, $Distort_k = 17$
- $Comb2$: $R_\alpha = 0.0420$, $G_\alpha = -0.5085$, $B_\alpha = 0.3695$, $H_\alpha = -0.0570$, $S_\alpha = -0.1978$, $V_\alpha = -0.4526$, $Blur_\sigma = 27$, $Noise_\sigma = 7$, $Distort_k = 68$
- $Comb3$: $R_\alpha = 0.1774$, $G_\alpha = -0.1150$, $B_\alpha = 0.1299$, $H_\alpha = -0.0022$, $S_\alpha = -0.2119$, $V_\alpha = -0.0747$, $Blur_\sigma = 1$, $Noise_\sigma = 6$, $Distort_k = 86$
- $Comb4$: $R_\alpha = -0.2599$, $G_\alpha = -0.0166$, $B_\alpha = -0.2702$, $H_\alpha = -0.4273$, $S_\alpha = 0.0238$, $V_\alpha = -0.2321$, $Blur_\sigma = 5$, $Noise_\sigma = 8$, $Distort_k = 8$
- $Comb5$: $R_\alpha = -0.2047$, $G_\alpha = 0.0333$, $B_\alpha = 0.3342$, $H_\alpha = -0.4400$, $S_\alpha = 0.2513$, $V_\alpha = 0.0013$, $Blur_\sigma = 35$, $Noise_\sigma = 6$, $Distort_k = 1$
- $Comb6$: $R_\alpha = -0.6613$, $G_\alpha = -0.0191$, $B_\alpha = 0.3842$, $H_\alpha = 0.3568$, $S_\alpha = 0.5522$, $V_\alpha = 0.0998$, $Blur_\sigma = 21$, $Noise_\sigma = 3$, $Distort_k = 37$

The datasets $Comb1$ through $Comb6$ are generated by randomly sampling parameters of each perturbation, e.g. blur, noise, distortion, and RGB and HSV channels, and combining these perturbations together. The parameters listed here are the parameters for the corresponding examples used in the experiment.

A.4 FID-MA AND L2D-MA DIFF

We illustrate the relationship between FID and Mean Accuracy (MA) Difference in Figure 9, and the relationship between L2 norm distance (L2D) and Mean Accuracy (MA) Difference in Figure 10. As shown in the figure, the FID space can better capture the difference among various factors affecting image quality better than the L2D space, i.e., the visual difference in the relationship trend for each factor in Figure 9 is more apparent than that in Figure 10.

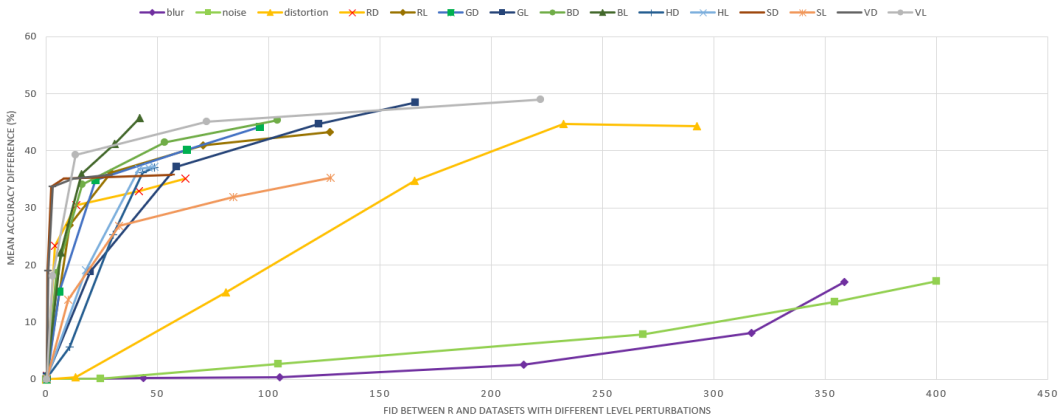

Figure 9: The relationship between FID and MA difference (in percentages).

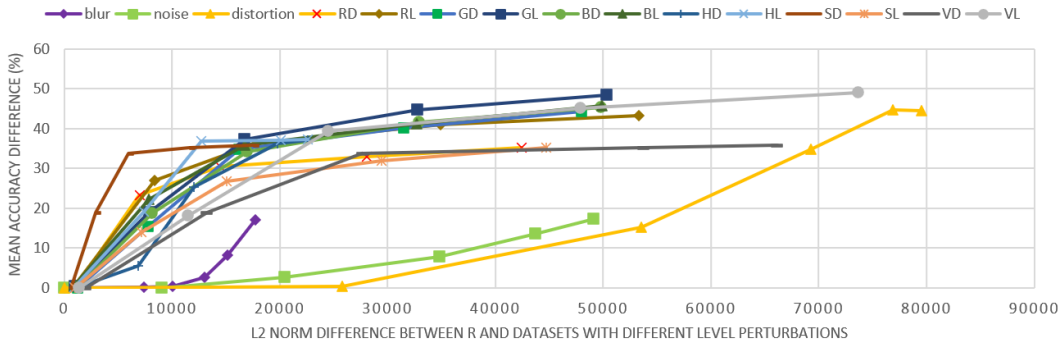

Figure 10: The relationship between L2 norm distance and MA difference (in percentages).

A.5 DATASET SETUP FOR HONDA AND AUDI

For Honda dataset, which contains more than 100 videos, we first select 30 videos that is most suitable for learning to steer task, then we extract 11,000 images from them at 1 FPS, and align them with the steering labels.

For Audi dataset, we use the "Gaimersheim" package which contains about 15,000 images with about 30 FPS, and then reduce to 15 FPS, keep about 7,500 images and align them with steering labels.

Both of them are then randomly splited into training/validation/test data with approximate ratio 20:1:2.

A.6 EXPERIMENT DATA

The tables shown in this section are a more fine-grained representation of our results, with metrics shown for each pairwise factor and level across methods. Section A.6.1 shows the data for mean accuracy measurements of each experiment in detail, and section A.6.2 shows the data for mean corruption error calculations. There are three main scenarios in these experiments: (1) single factor

scenario, where the model is trained on baseline data and tested on data corrupted by a single factor, (2) combination factor scenario, where the model is trained on baseline data and tested on data corrupted by a combination of single-factor perturbations, and (3) unseen factor scenario, where the model is trained on baseline data and tested on data corrupted by previously unseen factors. Single factor perturbations include blur, noise, distortion, and RGB and HSV channels on intensity levels L1 through L5. The parameters for these combinations are discussed in section A.3.

A.6.1 MEAN ACCURACY DATA

To quantify our results, we collected mean accuracy (MA) measurements from each experiment, for each pairwise factor and level across methods. Table 4 shows the mean accuracy measurements for blur, noise, and distortion factors. The same is of table 5, where mean accuracy is measured across levels of RGB or HSV color channels, where each channel serves as a single corruption factor. Table 6 presents the MA measurements for scenarios with a combination of factors, and Table 7 presents the MA measurements for scenes with previously unseen factors.

Method	Factor	L1	L2	L3	L4	L5
baseline	blur	88.2	88.1	86.1	81.2	73.3
	noise	88.3	86.0	81.4	76.4	73.2
	distortion	88.6	75.0	57.7	48.8	49.2
ours	blur	89.2	89.5	88.8	82.4	75.5
	noise	89.1	88.7	88.5	85.5	82.7
	distortion	89.1	85.5	63.1	56.5	50.6

Table 4: Mean accuracy of training (in %) using the baseline model and ours, tested on datasets with different levels of blur, noise, and distortion. Levels range from L1 to L5.

Method	Factor	DL5	DL4	DL3	DL2	DL1	LL1	LL2	LL3	LL4	LL5
baseline	R	53.2	55.4	57.9	65.1	87.8	87.7	61.4	52.1	47.4	45.1
	G	44.2	48.2	53.5	73.0	88.5	87.9	69.6	51.2	43.7	40.0
	B	43.0	46.8	54.3	69.7	88.2	87.7	66.2	52.5	47.1	42.6
	H	51.3	52.1	63.1	82.8	88.1	88.2	69.3	51.5	51.3	51.2
	S	58.4	63.8	72.6	83.9	88.1	88.3	74.5	61.6	56.5	53.2
	V	52.6	53.2	54.6	69.4	88.5	88.4	70.4	49.1	43.2	39.4
ours	R	87.3	88.8	89.4	89.5	89.4	89.4	89.4	89.7	89.1	87.4
	G	88.4	89.3	89.7	89.6	89.4	89.3	89.4	89.5	89.3	88.9
	B	89.0	89.5	89.5	89.2	89.4	89.3	89.4	89.5	89.3	88.9
	H	88.7	88.4	89.1	88.5	89.2	89.2	89.1	88.4	87.8	88.7
	S	85.7	87.8	88.2	89.0	89.3	89.3	89.3	88.5	88.2	84.5
	V	61.9	80.6	86.8	89.7	89.3	89.3	89.1	81.4	74.5	77.7

Table 5: Mean accuracy (MA) of training (in %) using the baseline model and ours, tested on datasets with different levels of R, G, B, and H, S, V channel values. DL denotes "darker level", which indicates a level in the darker direction of the channel, while LL indicates "lighter level", which indicates the lighter direction, on levels 1 to 5.

Method	Comb1	Comb2	Comb3	Comb4	Comb5	Comb6
baseline	59.7	54.0	40.9	50.0	54.0	56.3
ours	71.3	61.1	65.6	83.3	85.6	54.5

Table 6: Mean accuracy (MA) of training (in %) using the baseline model and ours, tested on datasets with several perturbations combined together, including blur, noise, distortion, RGB, and HSV.

Method	Unseen Factors	L1	L2	L3	L4	L5
baseline	motion_blur	76.4	69.7	62.62	61.1	60.33
	zoom_blur	85.57	83.66	81.79	79.97	78.15
	pixelate	88.15	88.21	88.04	88.27	88.1
	jpeg_comp	88.42	88.01	87.41	85.39	82.17
	snow	62.77	50.68	54.94	55.45	55.33
	frost	55.80	52.14	51.67	51.67	51.22
	fog	58.72	55	52.44	50.8	48.12
ours	motion_blur	76.0	68.1	59.4	57.9	58.1
	zoom_blur	87.4	85.8	83.6	81.8	79.9
	pixelate	89.6	89.7	89.6	89.6	89.5
	jpeg_comp	89.5	89.5	89.6	89.2	89.4
	snow	86.9	56.2	66.9	75.8	74.6
	frost	84.9	81.5	79.2	79.3	77.6
	fog	77.6	73.2	67.2	63.4	57.9

Table 7: Mean accuracy (MA) of training (in %) using the baseline model and ours, tested on datasets with previously unseen perturbations at 5 different levels. These types of unseen perturbations do not appear in the training data, and include motion blur, zoom blur, pixelate, jpeg compression loss, snow, frost, and fog, on intensity levels L1 to L5.

A.6.2 Mean Corruption Error Data

Mean corruption error (mCE) is a metric that can be used to measure the robustness of a model. We calculated the mean corruption errors based on the method used in (Hendrycks & Dietterich, 2019). Our calculations are shown here, in order to compare mCE across other methods and ours. Table 8 shows mCE calculations for testing on single factors, table 9 shows calculations for testing on combined factors, and table 10 shows calculations for testing on unseen factors.

Method	mCE	Blur	Noise	Distortion	R	G	B	H	S	V
Data Aug	51.3	90.3	70.7	83.8	32.3	30.4	31.4	36.8	43.8	**42.1**
Adv Training	74.4	101.6	90.2	**70.2**	73.7	57.5	60.2	72.2	71.0	73.1
Ours	**49.5**	**89.9**	**69.1**	85.9	**28.6**	**26.8**	**26.6**	**32.3**	**40.2**	46.0

Table 8: Calculations of mean corruption error (mCE) in % on single-factor test examples, compared on data augmentation, adversarial training, and our method. Here, the model is trained on baseline data and tested on data perturbed with a single perturbation factor. Single factor perturbation include blur, noise, distortion, RGB and HSV channels.

Method	mCE	Combo1	Combo2	Combo3	Combo4	Combo5	Combo 6
Data Aug	75.8	68.7	109.7	76.8	**27.7**	**25.3**	146.8
Adv Training	86.8	109.9	87.4	69.7	73.3	81.3	99.3
Ours	**63.8**	**71.3**	**84.5**	**58.2**	33.4	31.3	**104.1**

Table 9: Calculations of mean corruption error (mCE) in % on test examples with a combination of factors on data augmentation, adversarial training, and our method. The combination factor test data is generated by combining single factor perturbations onto one another.

Method	mCE	motion	zoom	pixelate	jpeg_comp	snow	frost	fog
Data Aug	81.5	108.5	90.1	101.9	88.8	68.8	45.0	**67.5**
Adv Training	89.9	**60.8**	101.1	104.2	98.0	92.1	75.6	97.7
Ours	**76.2**	106.2	**89.8**	**87.8**	**76.9**	**63.2**	**41.1**	68.4

Table 10: Calculations of mean corruption error (mCE) in % on test examples with previously unseen factors on data augmentation, adversarial training, and our method. The unseen factors include motion blur, zoom blur, pixelate, jpeg compression, snow, frost, and fog.