FICGen: Frequency-Inspired Contextual Disentanglement for Layout-driven Degraded Image Generation

Supplementary Material

The supplementary material provides a more comprehensive evaluation of our proposed *FICGen* for degraded image synthesis, which is organized as follows:

- Section 1: Dataset Introductions.
- Section 2: Implementation Details.
- Section 3: Additional Experimental Results.
- Section 4: Generative Visualizations.
- Section 5: Limitations and Future Work.

Dataset	Train	Test	Total	Instances	Classes	Degradation mode
ExDARK [9]	5,145	2,218	7,363	23,710	12	low-light
RUOD [6]	9,800	4,200	14,000	74,904	10	underwater
DIOR [8]	5,862	11,738	17,600	192,472	20	aerial
DAWN [7]	5,544	410	5,954	43,869	5	adverse weather
VOC 2012 [5]	10,582	1,449	12,031	33,149	20	blur

Table 1. An overview of five degraded datasets.

1. Dataset Introductions

To comprehensively evaluate the generative effectiveness of our *FICGen* across diverse degraded conditions, we conduct extensive experiments on **five** widely used benchmarks: ExDARK [9], DIOR [8], RUOD [6], DAWN [7], and blurred VOC 2012 [5]. Tab. 1 summarizes the key statistics of each dataset, including the numbers of training and testing samples, total annotated instances, and degradation types. Two considerations are noteworthy:

(1). Due to the relatively small size of DAWN, we employ conventional data augmentation techniques—such as scaling, translation, and horizontal flipping—to expand the training set from 590 to 5,544 samples. Additionally, given the extreme scarcity of "bicycle" instances in DAWN, this category is omitted from the evaluation.

(2). Following [14], we generate blur kernels on the fly to construct the blurred VOC 2012 [5] dataset, which is used to inspect the adaptability of *FICGen* to mild degradations. Collectively, these datasets span a broad spectrum of degradation types, from severe low-light, underwater, aerial, and adverse weather conditions (*i.e.*, rain, fog, snow, sandstorms) to mild blur, thereby enabling a thorough evaluation of the generalizability and robustness of *FICGen* in real-world degraded scenarios.

2. Implementation Details

As mentioned in the main paper, our *FICGen* is built upon the pre-trained Stable Diffusion model (v1.5) [13] and is incorporated into the mid-level (8×8) feature layers and the lowest-resolution (16×16) decoder layers of the denoising U-Net. All images are processed at a fixed resolution of 512×512 during the training and inference phases. To evaluate its generative capacity for densely distributed instances in degraded contexts, we restrict each image to at most $N = 15$ objects, in line with the AeroGen setting [16]. For a fair comparison, all competing L2I methods (*i.e.*, MIGC [19], CC-Diff [18]) are trained under identical configurations (*i.e.*, learning rate, training epochs, and the number of instances per image).

For **fidelity evaluation**, the same number of real and synthetic images are resized to a fixed resolution of 512×512, we then employ the library "torch-fidelity" [11] to compute the FID scores based on a pre-trained Inception-V3 model. It should be noted that, to ensure a fair comparison with AeroGen (CVPR 2025), the FID scores for remote sensing scenarios are computed using an Inception-V3 model fine-tuned on the RSICD [10] dataset, following the procedure in [17]. For **alignment evaluation**, we utilize pre-trained downstream detectors, *i.e.*, Faster R-CNN [12] and Cascade R-CNN [1], to predict detection results, which are compared against ground-truth bounding boxes and classes to compute alignment metrics. For **trainability evaluation**, we construct a synthetic training set equal in size to the real dataset, which serves as an auxiliary resource to enhance downstream detectors. To be specific, ground-truth bounding boxes (bboxes) from the training split of the degraded datasets are used as layout conditions for generating synthetic images. Following [3], we first discard bboxes smaller than 0.2% of the image area, then apply data augmentation by randomly flipping bboxes with a probability of 0.8 and shifting them within 128 pixels. The generated synthetic images are combined with the real ones across various training settings. All experiments follow the default training and testing protocols of MMDetection 2.25.3 [2], with all images uniformly resized to 512×512 and trained under a standard $1\times$ schedule.

Detector	Method	FID↓	Object Detection (AP) for Sampled Classes /%						mAP↑	AP_{50} ↑	AP_{75} ↑
DIOR-H [8]			windmill	airport	stadium	ballfield	airplane	golffield			
Faster R-CNN	Oracle	-	29.0	32.2	26.7	50.2	33.5	34.7	33.4	55.6	35.0
Faster R-CNN	MIGC [19]	31.64	3.4	33.6	18.0	32.4	6.8	43.8	21.8	38.4	17.5
Faster R-CNN	CC-Diff [18]	30.88	5.6	31.4	21.4	44.0	9.5	50.6	23.6	42.4	21.4
Faster R-CNN	FICGen (ours)	31.25	16.0	33.5	16.9	50.0	23.8	51.2	27.6	48.7	27.6
RUOD [6]			starfish	echinus	fish	scallop	corals	turtle			
Faster R-CNN	Oracle	-	45.9	44.1	41.9	37.5	45.9	67.8	50.5	80.2	54.4
Faster R-CNN	MIGC [19]	26.50	20.6	12.2	14.9	13.5	22.8	48.2	27.2	54.1	24.6
Faster R-CNN	CC-Diff [18]	25.21	22.7	13.3	16.8	12.5	25.6	50.7	29.7	58.4	27.9
Faster R-CNN	FICGen (ours)	25.10	32.1	28.5	25.0	23.7	33.7	52.5	37.0	68.6	36.5
blurred VOC 2012 [5]			aeroplane	boat	cow	bottle	sheep	person			
Faster R-CNN	Oracle	-	43.7	26.9	24.3	19.7	27.9	34.5	31.5	56.5	32.4
Faster R-CNN	MIGC [19]	62.66	42.4	27.0	23.9	20.2	21.8	30.3	34.7	65.8	33.1
Faster R-CNN	CC-Diff [18]	62.20	44.0	28.6	26.5	19.6	27.1	32.5	36.7	67.6	36.3
Faster R-CNN	FICGen (ours)	58.02	49.0	36.1	36.2	23.2	35.4	35.4	40.7	70.3	42.7
VOC 2012 [5]			aeroplane	boat	cow	bottle	sheep	person			
Faster R-CNN	Oracle	-	57.1	42.9	48.7	33.2	46.1	48.1	48.3	76.8	52.5
Faster R-CNN	AeroGen [16]	45.21	38.1	27.7	30.9	28.7	35.4	29.5	36.8	65.7	36.4
Faster R-CNN	MIGC [19]	50.60	47.3	33.9	44.9	24.1	35.0	34.3	45.1	78.5	47.4
Faster R-CNN	CC-Diff [18]	48.70	55.0	36.2	46.4	27.2	37.8	35.5	47.9	79.1	52.0
Faster R-CNN	FICGen (ours)	48.93	57.7	47.9	49.9	37.4	45.9	44.6	54.2	83.5	60.2
ExDARK [9]			bicycle	motorbike	boat	car	cup	bottle			
Cascade R-CNN	Oracle	-	48.5	33.6	30.4	40.2	29.0	32.1	37.2	65.8	37.8
Cascade R-CNN	MIGC [19]	45.76	45.9	25.4	26.6	28.8	21.9	21.5	32.4	63.5	29.5
Cascade R-CNN	CC-Diff [18]	44.26	48.5	32.5	30.6	33.0	23.3	25.9	35.1	65.6	34.1
Cascade R-CNN	FICGen (ours)	42.40	57.3	38.5	29.4	43.9	35.8	34.8	42.5	73.0	45.1
DAWN [7]			motorcycle	person	bus	truck	car	-			
Cascade R-CNN	Oracle	-	19.4	20.6	27.0	26.7	42.8	-	27.3	46.4	26.3
Cascade R-CNN	MIGC [19]	70.10	9.3	7.2	17.0	14.6	14.8	-	12.6	32.3	8.6
Cascade R-CNN	CC-Diff [18]	68.56	13.2	9.4	19.4	17.6	19.0	-	15.7	33.9	14.8
Cascade R-CNN	FICGen (ours)	68.31	25.3	20.3	19.8	19.6	28.2	-	22.6	44.3	21.5

Table 2. Quantitative comparison of generative fidelity and alignment on five degraded datasets (DIOR-H [8], ExDARK [9], RUOD [6], DAWN [7], and blurred VOC 2012 [5]) and one natural image dataset (VOC 2012 [5]). The performance is evaluated using off-the-shelf detectors Faster R-CNN (R50) [12] and Cascade R-CNN (R50) [1] on synthetic test images generated by different L2I methods. "*Oracle*" denotes the real test set baseline (*i.e.*, upper bound). The top-2 performers are marked in red and underlined.

3. Additional Experimental Results

3.1. More Quantitative Evaluations

Alignment. Tab. 2 reports a comprehensive evaluation of generative fidelity (FID) and alignment (detection AP) across **five** degraded benchmarks (DIOR-H [8], ExDARK [9], RUOD [6], DAWN [7], and blurred VOC 2012 [5]) and one natural benchmark (VOC 2012 [5]). The alignment is assessed using two pre-trained detectors, Faster R-CNN (R50) [12] and Cascade R-CNN (R50) [1], enabling a fair comparison with prior L2I approaches. Overall, our FICGen consistently delivers superior alignment and fidelity, attaining a **27.6 mAP** on DIOR-H, outperforming MIGC (21.8) and CC-Diff (23.6) by **5.8** and **4.0** points, respectively. The gains are particularly pronounced on challenging semantic categories, such as "*windmill*" (16.0 *vs.* 3.4/5.6) and "*airplane*" (23.8 *vs.* 6.8/9.5).

Similar improvements are observed across other benchmarks. On RUOD, which involves dense underwater instances, FICGen achieves a **37.0 mAP**, surpassing CC-Diff by **7.3** and MIGC by **9.8**, owing to its frequency-inspired contextual disentanglement that mitigates the submersion of underwater objects against homogeneous aquatic backgrounds. On blurred VOC 2012, our method elevates AP for almost all categories such as "*cow*" (36.2 *vs.* 23.9/26.5) and "*sheep*" (35.4 *vs.* 21.8/27.1), resulting in a **40.7 mAP**, which is a clear lead over CC-Diff (36.7) and MIGC (34.7). For ExDARK and DAWN, which represent extreme low-light and adverse-weather conditions, FICGen delivers notable boosts in categories with attenuated high-frequency cues, such as "*motorbike*" (38.5 *vs.* 25.4/32.5) and "*motorcycle*" (25.3 *vs.* 9.3/13.2), achieving **42.5** and **22.6 mAP**, respectively.

Crucially, these improvements can be attributed to FICGen's contextual disentanglement mechanism, which effectively separates high-frequency instance details from the dominant low-frequency surroundings, while simultaneously preserving essential degraded contextual characteristics such as illumination and texture.

Trainability. Following the trainability protocol in [3], we leverage ground-truth bounding boxes from the degraded

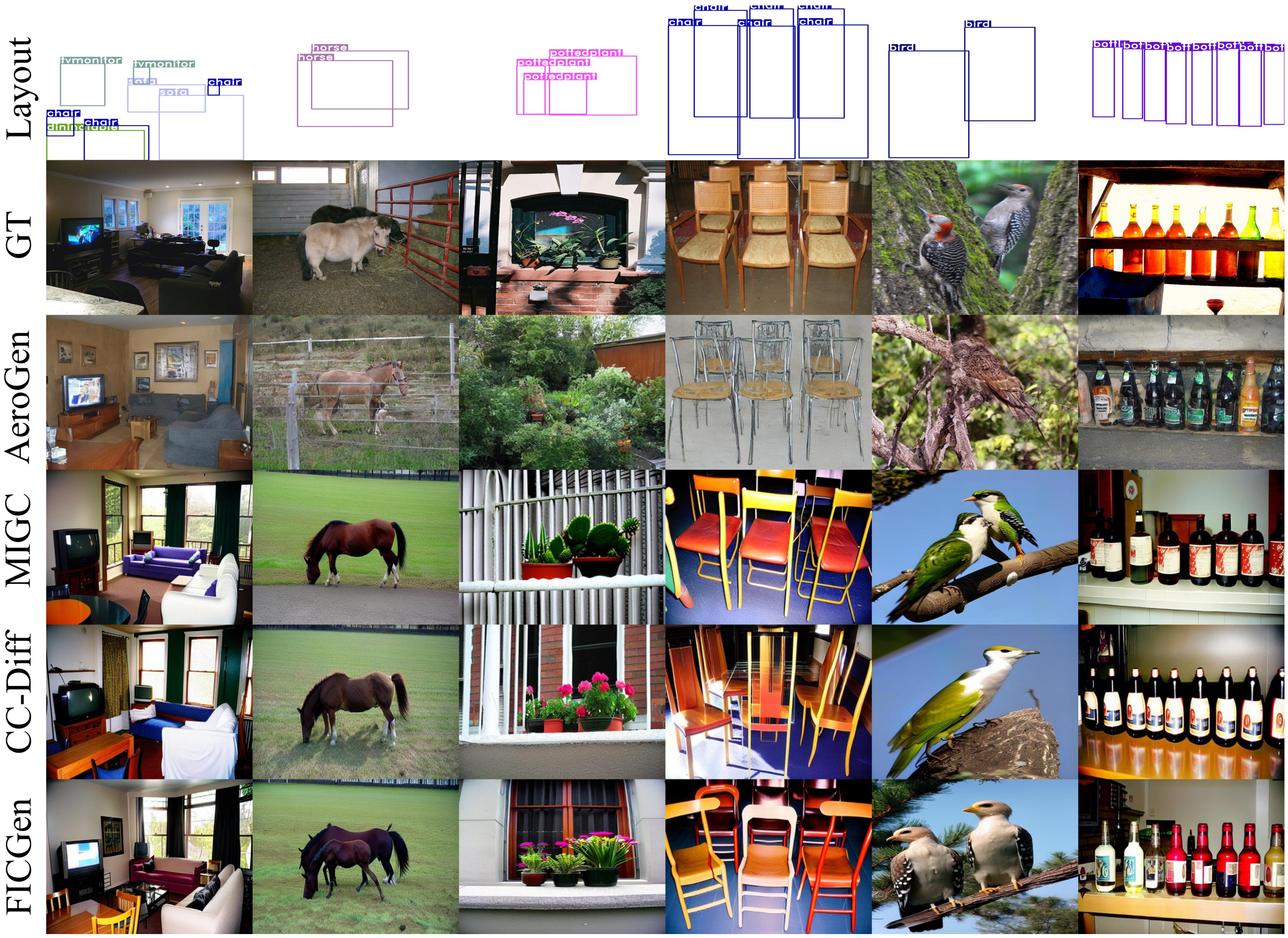

Figure 1. Qualitative comparison of natural images (VOC 2012 [5]) generated by different L2I methods. Zoom in for more detail.

datasets as layout inputs to synthesize additional training images, which are combined with real samples to effectively double the training set size. As summarized in Tabs. 3– 5, our proposed FICGen consistently delivers superior gains in downstream detection performance across both degraded and natural scenarios. On the ExDARK benchmark, FICGen achieves the highest mean AP, with notable improvements for semantic categories such as "*motorbike*" (**+3.8**) relative to the Oracle upper bound and "*people*" (**+1.6**) over the closest CC-Diff. On DAWN, despite the small scale of available data, FICGen yields competitive boosts, including **+4.9** for "*motorcycle*" and **+4.4** for "*bus*." More importantly, on natural VOC 2012, which reflects the generalization capability, FICGen achieves the best overall mean AP (**50.5**), surpassing CC-Diff by **+0.9** and MIGC by **+1.1**. These downstream trainability results confirm that the degraded images synthesized by FICGen not only alleviate the scarcity of training data in adverse conditions but also serve as effective auxiliary resources to enhance the accuracy of downstream detectors.

Robust Control. To further investigate the generative robustness of FICGen towards spatially entangled instances under degraded conditions, Tab. 6 presents a detailed alignment comparison on four benchmarks with varying occlusion levels. Each dataset is divided into three occlusion levels, *i.e.*, Sparse, Partial, and Heavy, according to the number of instances and their mutual occlusion measured by Intersection over Union. A pre-trained Faster R-CNN (R50) is then used to assess the alignment between the generated instances and their corresponding layouts at each occlusion level.

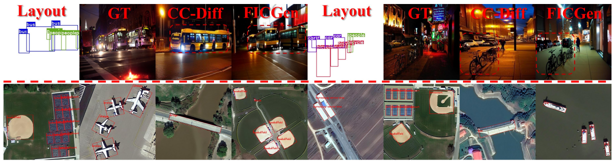

Figure 2. Further qualitative comparison with CC-Diff on ExDARK and generalization to oriented layout inputs [4].

As reported in Tab. 6, our FICGen demonstrates remarkable robustness across varying occlusion levels, significantly

category	Oracle	MIGC [19]	CC-Diff [18]	FICGen	category	Oracle	MIGC [19]	CC-Diff [18]	FICGen
bicycle	46.4	48.3	49.6	47.8	chair	25.4	26.8	26.8	27.5
boat	30.2	31.1	29.7	31.7	cup	28.6	28.2	28.7	29.6
bottle	31.8	31.6	31.1	32.8	dog	44.1	47.9	47.6	48.5
bus	56.8	60.8	59.8	59.9	motorbike	33.3	34.4	34.1	37.1
car	38.1	38.3	38.4	38.8	people	32.9	32.5	32.4	34.0
cat	41.8	41.4	42.0	42.4	table	20.6	24.1	24.1	22.5
					All (**mAP %**)	35.8	37.1	37.0	37.7

Table 3. Detection accuracy (%) per class on the ExDARK *test* set using Faster R-CNN (R50) [12], trained on 5.1k real and 5.1k synthetic images generated by different L2I methods.

category	Oracle	MIGC [19]	CC-Diff [18]	FICGen
motorcycle	18.0	23.0	20.0	22.9
person	19.6	19.5	19.5	20.2
bus	21.8	21.6	27.8	26.2
truck	23.4	19.6	21.6	20.5
car	41.1	39.4	40.0	39.6
All (**mAP %**)	24.8	24.6	25.8	25.9

Table 4. Detection accuracy (%) per class on the DAWN *test* set using Faster R-CNN (R50) [12], trained on 5.5k real and 5.5k synthetic images generated by different L2I methods.

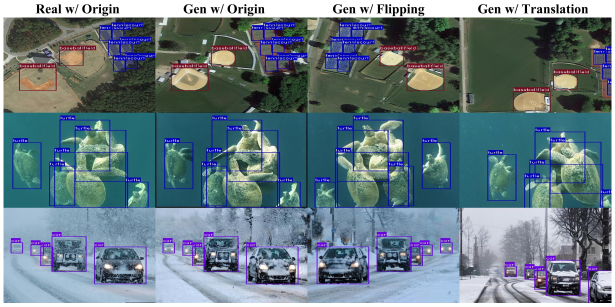

Figure 3. Visualization results generated by FICGen conditioned on augmented geometric layouts.

outperforming prior L2I approaches. On sparsely occluded cases, where inter-instance interference is minimal, FICGen already yields clear gains, achieving a mAP of **44.8** on DIOR-H and **63.8** on ExDARK, surpassing CC-Diff by **6.1** and **3.2**, respectively. These improvements become more pronounced under partial occlusion, where overlapping instances introduce complex spatial entanglement. For example, on RUOD, FICGen attains a **46.5 mAP**, outperforming MIGC and CC-Diff by **16.1** and **8.3**, respectively. The benefits are most substantial under heavy occlusion, where conventional methods often suffer from "object omission and merging" due to severe spatial collisions. On DAWN and RUOD, which contain up to **55**% and **59**% heavily occluded instances, FICGen secures **20.0** and **33.5 mAP**, nearly doubling the performance of MIGC (**11.1** and **17.8**) and substantially outperforming CC-Diff (**13.2** and **26.3**). Collectively, these results underline FICGen's ability to faithfully adhere to user-specified layouts and mitigate occlusion-induced distortions, thereby generating structurally consistent degraded images that benefit downstream perception tasks.

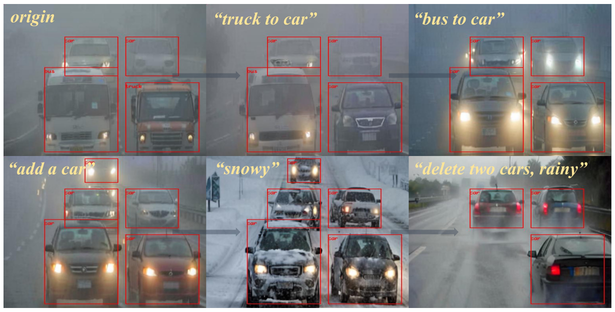

Figure 4. Visualization results of the continuous generation of instance interactivity (*i.e*., addition, removal, transform and weather change) by our FICGen on the DAWN dataset.

Inference Efficiency Analysis. Tab. 7 reports the number of trainable parameters and the per-image inference time for identical settings on the DIOR dataset. FICGen contains only around a third of the parameters of AeroGen (∼304M *vs.* ∼905M). Compared with CC-Diff, FICGen achieves the highest mAP while maintaining a comparable overhead, requiring only an additional 40M trainable parameters and incurring a modest 2-second inference delay.

3.2. More Qualitative Evaluations

Fig. 1 presents a qualitative comparison of generation results produced by different L2I methods on natural scenes from VOC 2012. Existing approaches exhibit notable limitations in generative quality, often suffering from hallucination artifacts such as missing objects and incorrect merging. For instance, AeroGen generates synthesized images whose visual characteristics deviate substantially from real-world counterparts, while MIGC and CC-Diff frequently merge adjacent objects, such as **two horses being fused into one (second column) or three potted plants collapsing into two or overflowing excessively (third column)**. Moreover, when

category	Oracle	MIGC [19]	CC-Diff [18]	FICGen	category	Oracle	MIGC [19]	CC-Diff [18]	FICGen
aeroplane	57.1	58.0	59.6	59.2	diningtable	42.8	44.5	42.3	45.2
bicycle	51.6	54.3	53.1	53.6	dog	59.6	59.8	60.9	61.9
bird	52.2	53.6	54.1	55.7	horse	54.9	55.7	55.2	55.4
boat	42.9	42.5	44.4	47.3	motorbike	56.2	58.7	59.1	57.8
bottle	33.2	34.7	36.3	34.9	person	48.1	47.9	48.1	48.9
bus	59.4	60.1	60.1	62.2	pottedplant	26.9	25.8	26.0	27.9
car	43.0	43.6	44.6	45.9	sheep	46.1	46.7	45.7	47.3
cat	62.2	61.5	62.1	63.0	sofa	43.6	47.9	47.5	47.6
chair	26.5	28.7	28.9	29.6	train	59.9	61.4	62.0	61.6
cow	48.7	51.6	51.0	51.0	tvmonitor	51.4	50.9	50.7	53.1
					All (**mAP %**)	48.3	49.4	49.6	50.5

Table 5. Detection accuracy (%) per class on natural VOC 2012 *test* set using Faster R-CNN (R50) [12], trained on 10.6k real and 10.6k synthetic images generated by different L2I methods.

Dataset	Method	Sparse (33%/39%/23%/27%)			Partial (21%/15%/18%/18%)			Heavy (46%/46%/59%/55%)		
		mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}
	Oracle	45.9	70.2	50.9	35.2	53.3	38.6	28.7	47.6	30.7
DIOR-H [8]	MIGC [19]	33.1	64.4	30.7	19.1	38.8	16.7	16.1	33.9	12.9
	CC-Diff [18]	38.7	66.3	40.7	24.4	42.9	25.0	19.5	36.5	19.1
	FICGen	44.8	72.4	48.8	30.1	49.2	32.5	24.1	42.6	24.9
	Oracle	55.6	87.4	65.7	42.6	73.4	46.3	28.7	58.6	25.3
ExDARK [9]	MIGC [19]	59.3	95.1	68.0	44.7	85.8	43.1	24.9	58.5	16.7
	CC-Diff [18]	60.6	95.1	72.2	47.7	84.3	49.0	26.1	58.5	19.3
	FICGen	63.8	95.3	75.8	54.0	89.0	59.8	32.1	65.3	28.0
	Oracle	56.2	81.2	64.3	53.0	81.6	58.5	48.1	79.6	51.0
RUOD [6]	MIGC [19]	40.1	75.2	39.1	30.4	63.7	25.8	17.8	39.5	13.4
	CC-Diff [18]	47.8	80.6	51.3	38.2	73.1	38.3	26.3	54.2	23.0
	FICGen	54.3	85.6	62.2	46.5	80.4	49.6	33.5	65.3	31.1
	Oracle	36.1	56.5	35.2	34.9	56.3	42.9	23.8	45.3	21.4
DAWN [7]	MIGC [19]	25.3	52.3	21.3	17.8	38.8	13.0	11.1	29.9	5.7
	CC-Diff [18]	38.2	76.7	33.6	26.4	50.9	26.3	13.2	32.1	9.2
	FICGen	44.5	71.7	48.1	34.5	54.0	40.7	20.0	40.8	16.1

Table 6. Quantitative comparison of alignment across four datasets under three occlusion degrees, with detection evaluated by Faster R-CNN (R50). The numbers in brackets represent the proportion of each occlusion level within the dataset.

Method	Trainable Params (M)	Inf.Time (s/img)	Inf. Mem (GB)	mAP
CC-Diff	∼ 261	8	∼ 12.2	26.4
AeroGen	∼ 905	8	∼ 13.9	29.8
FICGen (ours)	∼ 304	10	∼ 12.6	31.2

Table 7. Comparison of trainable parameters and inference efficiency with AeroGen [16] and CC-Diff [18].

processing densely distributed objects, such as the array of bottles in column six, these methods often fail to preserve accurate spatial arrangement and object counts. In contrast, although primarily designed for degraded image generation, our FICGen demonstrates superior performance in natural scenarios, particularly in maintaining object quantity, spatial positioning, and scale. Specifically, it accurately reproduces **highly overlapping objects, including the two horses in the second column and six closely packed chairs in the fourth column, while preserving fidelity for small-scale targets, exemplified by the chair in the first column.**

In the additional low-light results shown in Fig. 2, CC-Diff **suffers from the illusion of merging three "*buses*" into one and omitting dense "*motorbike*" instances**, whereas our FICGen effectively resolves such contextual illusion issues. What's more, as shown in the second row of Fig. 2, FICGen exhibits strong adaptability to oriented layout controls, ensuring consistent generation quality under such constraints.

Figs. 3 and 4 present the generative results on layout manipulation. Fig. 3 demonstrates that FICGen maintains robust generation for augmented layouts, such as flipping and translation, while Fig. 4 shows that FICGen consistently produces coherent content under continuous layout modifications, including object addition (*i.e.*, "*add a car*"),

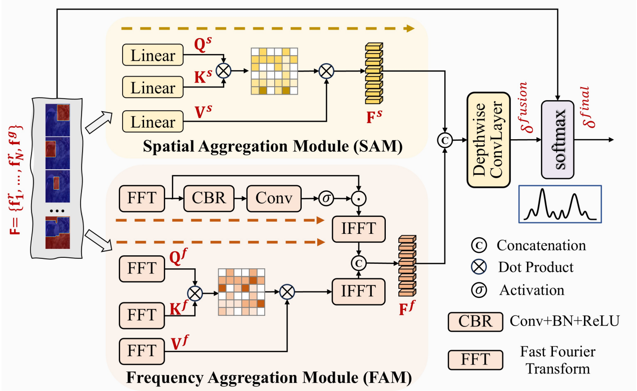

Figure 5. Architectural details of the proposed Adaptive Spatial-Frequency Aggregation Module (ASFA).

Fusion	airplane	mAP	AP_{50}	AP_{75}
Depth-wise	23.8	27.6	48.7	27.6
Point-wise	16.0	26.4	48.1	26.4

Table 8. The ablation results of different fusion strategies for SAM and FAM outputs.

removal (*i.e.*, "*delete two cars*"), categorical transformation (*i.e.*, "*truck to car*"), and weather changes (*i.e.*, "*snowy, rainy*"), highlighting its flexibility and user-controllable interactivity. These qualitative results further underscore the advantages of FICGen in degraded image generation.

Fig. 6 presents the detection results of Faster R-CNN *wi/wo* FICGen for auxiliary training. The synthesized degraded images from FICGen notably improve the detector's localization accuracy, particularly for remote sensing instances such as "*baseballfield*" and "*people*" in low-light conditions.

3.3. More Architecture Details

Fig. 5 illustrates the architectural details of the proposed Adaptive Spatial-Frequency Aggregation (ASFA) module. Inspired by [15], we adopt a dual-branch spatial-frequency aggregation strategy to integrate the disentangled degraded representations in the latent space. The spatial branch captures contextual dependencies, including semantic correlations and degradation similarities among various objects. Concurrently, the frequency branch focuses on fine-grained attributes such as edge structures and texture details. Next, we further fuse the dual-stream outputs of the SAM and FAM at a lower cost by using a single-layer depthwise separable convolution to enhance local perception within different degraded regions. Finally, adaptive weights for context-aware aggregation are obtained via a softmax operation. As shown in Tab. 8, this fusion strategy substantially outperforms the point-wise alternative, particularly for small objects like "*airplane*", where AP improves from 16.0 to 23.8. These results highlight the effectiveness of our adaptive aggregation in capturing both global dependencies and local structural cues under complex degradation scenarios.

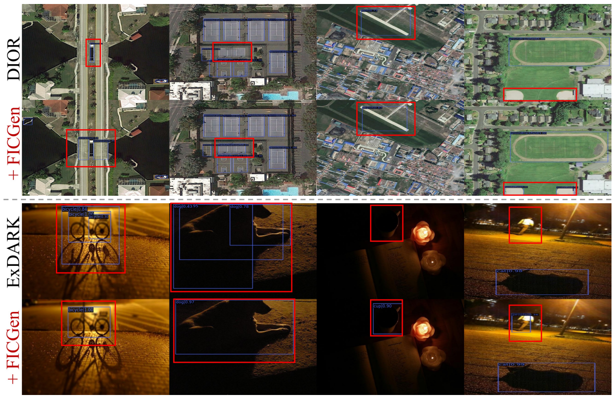

Figure 6. Detection results of Faster R-CNN (R50) trained on real images *vs.* real & FICGen-generated images.

4. Generative Visualizations

Figs. 8– 12 further showcase the controllable generation capabilities of FICGen, highlighting its ability to address a diverse range of degraded environments, including mild blur, low illumination, underwater, remote sensing, and severe adverse weather conditions. These visualizations emphasize the adaptability and robustness of FICGen in accurately representing the distinctive visual and semantic characteristics of each degraded context.

In particular, Fig. 8 presents synthetic results on the blurred VOC 2012 dataset, illustrating FICGen's capability to handle mild degradations while preserving semantic integrity for categories such as "*bird*", "*cat*", and "*sheep*", without introducing noticeable artifacts or structural inconsistencies. Moreover, even under moderate motion blur, FICGen successfully renders distinguishable shapes for objects like "*train*" and "*horse*", thereby retaining perceptible fine-grained blurred details for both foreground objects and surrounding backgrounds.

Figs. 9 and 10 present that our FICGen excels in replicating the severe conditions of low-illumination and underwater scenarios, while preserving the realism of complex lighting and distortion effects, ultimately delivering semantically aligned and visually realistic degraded samples. Fig. 11 further showcases synthetic results on remote sensing scenes, underscoring FICGen's capability to generate dense object clusters while preserving coherent spatial relationships between foreground instances and their surroundings. In particular, FICGen accurately renders densely distributed targets such as "*ship*" in port areas, while maintaining the correct number of objects. Moreover, it captures contextual con-

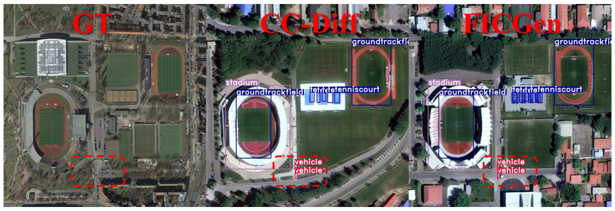

Figure 7. Failure cases of FICGen, where red dashed boxes denote the missing instances.

sistency by generating "*vehicles*" precisely aligned along "*overpasses*", ensuring that the synthesized objects seamlessly integrate with the underlying scene structure.

Fig. 12 presents synthetic samples generated by FICGen under four adverse weather conditions—fog, snow, sand, and rain—on the DAWN dataset, illustrating its strong generalization across visually diverse and challenging degraded scenarios. We can see that FICGen effectively captures the distinctive visual properties of each weather type: *i.e*., the diffuse scattering and visibility attenuation of "*fog*". Beyond instance-level fidelity, FICGen demonstrates the capability to restore critical scene-level elements, such as lane markings and road boundaries, even when they are partially occluded or obscured by rain or sand. Notably, in the "*sand*" condition, FICGen generates a highly realistic visual atmosphere, reproducing the chromatic desaturation observed in authentic sandstorm scenarios.

5. Limitations and Future Work

Fig. 7 illustrates the failure case, where FICGen struggles to synthesize high-resolution remote sensing images with precise representations of small objects such as "*vehicles*." This limitation primarily arises from the inherent downsampling operations in latent diffusion models, which suppress fine-grained structural details. Future work will focus on extending the frequency-inspired paradigm to 3D content generation, such as camouflaged video synthesis, and exploring richer control modalities beyond bounding boxes, including semantic masks, to enhance the precision and controllability of contextual generation in degraded scenarios.

References

[1] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 1, 2

[2] Kai Chen, Jiaqi Wang, Jiangmiao Pang, and et al. Mmdetection: Openmmlab detection toolbox and benchmark, 2019. 1

[3] Kai Chen, Enze Xie, Zhe Chen, Yibo Wang, Lanqing Hong, Zhenguo Li, and Dit-Yan Yeung. Geodiffusion: Text-prompted geometric control for object detection data generation. *arXiv preprint arXiv:2306.04607*, 2023. 1, 2

[4] Gong Cheng, Jiabao Wang, Ke Li, Xingxing Xie, Chunbo Lang, Yanqing Yao, and Junwei Han. Anchor-free oriented proposal generator for object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022. 3

[5] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 1, 2, 3

[6] Chenping Fu, Risheng Liu, Xin Fan, Puyang Chen, Hao Fu, Wanqi Yuan, Ming Zhu, and Zhongxuan Luo. Rethinking general underwater object detection: Datasets, challenges, and solutions. *Neurocomputing*, 517:243–256, 2023. 1, 2, 5, 9

[7] Mourad A Kenk and Mahmoud Hassaballah. Dawn: vehicle detection in adverse weather nature dataset. *arXiv preprint arXiv:2008.05402*, 2020. 1, 2, 5, 10

[8] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159:296–307, 2020. 1, 2, 5, 9

[9] Yuen Peng Loh and Chee Seng Chan. Getting to know low-light images with the exclusively dark dataset. *Computer Vision and Image Understanding*, 178:30–42, 2019. 1, 2, 5, 8

[10] Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2183–2195. 1

[11] Anton Obukhov, Maximilian Seitzer, Po-Wei Wu, Semen Zhydenko, Jonathan Kyl, and Elvis Yu-Jing Lin. High-fidelity performance metrics for generative models in pytorch, 2020. Version: 0.3.0, DOI: 10.5281/zenodo.4957738. 1

[12] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1, 2, 4, 5

[13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[14] Mohamed Sayed and Gabriel Brostow. Improved handling of motion blur in online object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1706–1716, 2021. 1

[15] Yanguang Sun, Chunyan Xu, Jian Yang, Hanyu Xuan, and Lei Luo. Frequency-spatial entanglement learning for camouflaged object detection. In *European Conference on Computer Vision*, pages 343–360. Springer, 2024. 6

[16] Datao Tang, Xiangyong Cao, Xuan Wu, Jialin Li, Jing Yao, Xueru Bai, Dongsheng Jiang, Yin Li, and Deyu Meng. Aerogen: enhancing remote sensing object detection with diffusion-driven data generation. *arXiv preprint arXiv:2411.15497*, 2024. 1, 2, 5

[17] Yonghao Xu, Weikang Yu, Pedram Ghamisi, Michael Kopp, and Sepp Hochreiter. Txt2img-mhn: Remote sensing image generation from text using modern hopfield networks. *IEEE Transactions on Image Processing*, 32:5737–5750, 2023. 1

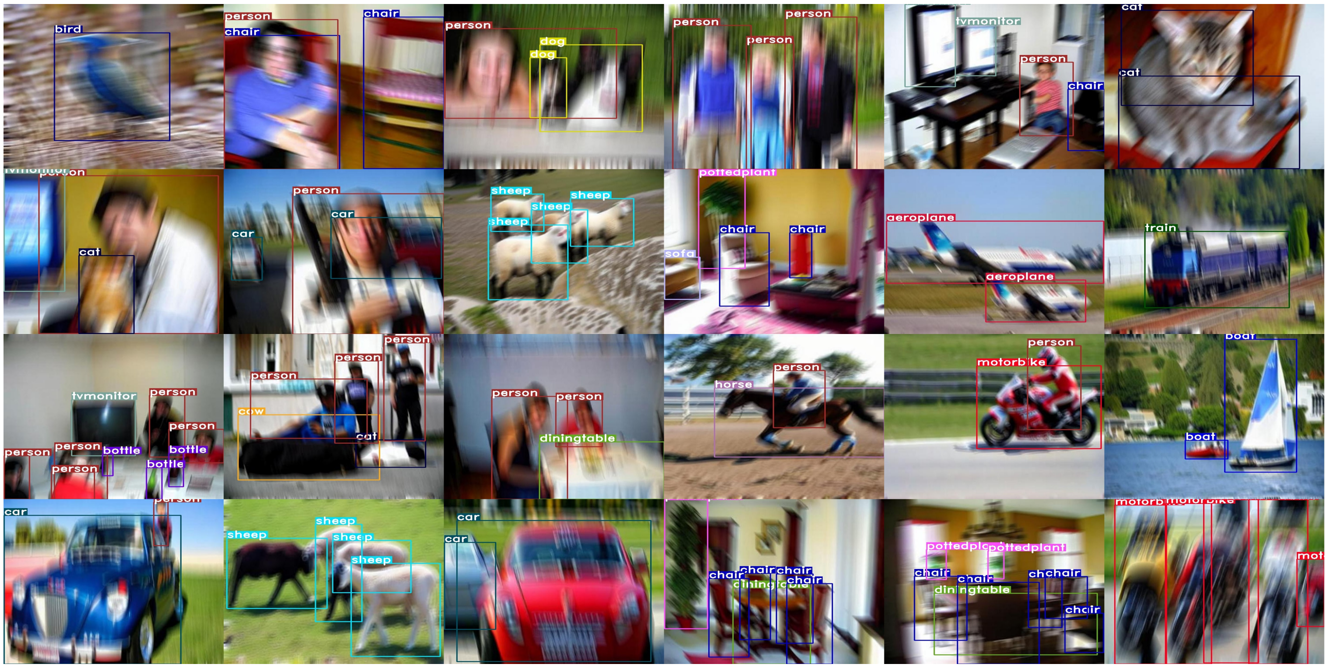

Figure 8. Visualization results generated by our FICGen under the mild blur (blurred VOC 2012), with geometric layouts and corresponding object categories superimposed on the generated images.

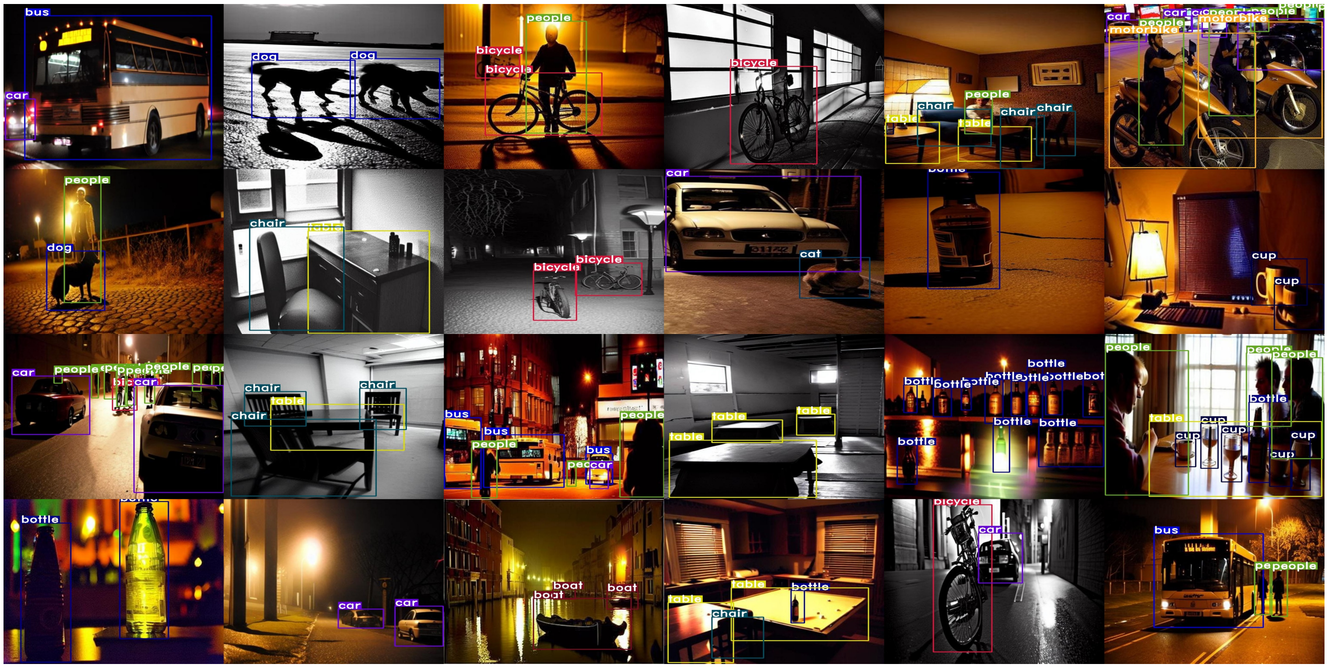

Figure 9. Visualization results generated by our FICGen under low-light conditions (ExDARK [9]), with geometric layouts and corresponding object categories superimposed on the generated images.

[18] Mu Zhang, Yunfan Liu, Yue Liu, Hongtian Yu, and Qixiang Ye. Cc-diff: Enhancing contextual coherence in remote sensing image synthesis. *arXiv preprint arXiv:2412.08464*, 2024. 1, 2, 4, 5

[19] Dewei Zhou, You Li, Fan Ma, Xiaoting Zhang, and Yi Yang. Migc: Multi-instance generation controller for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6818–6828,

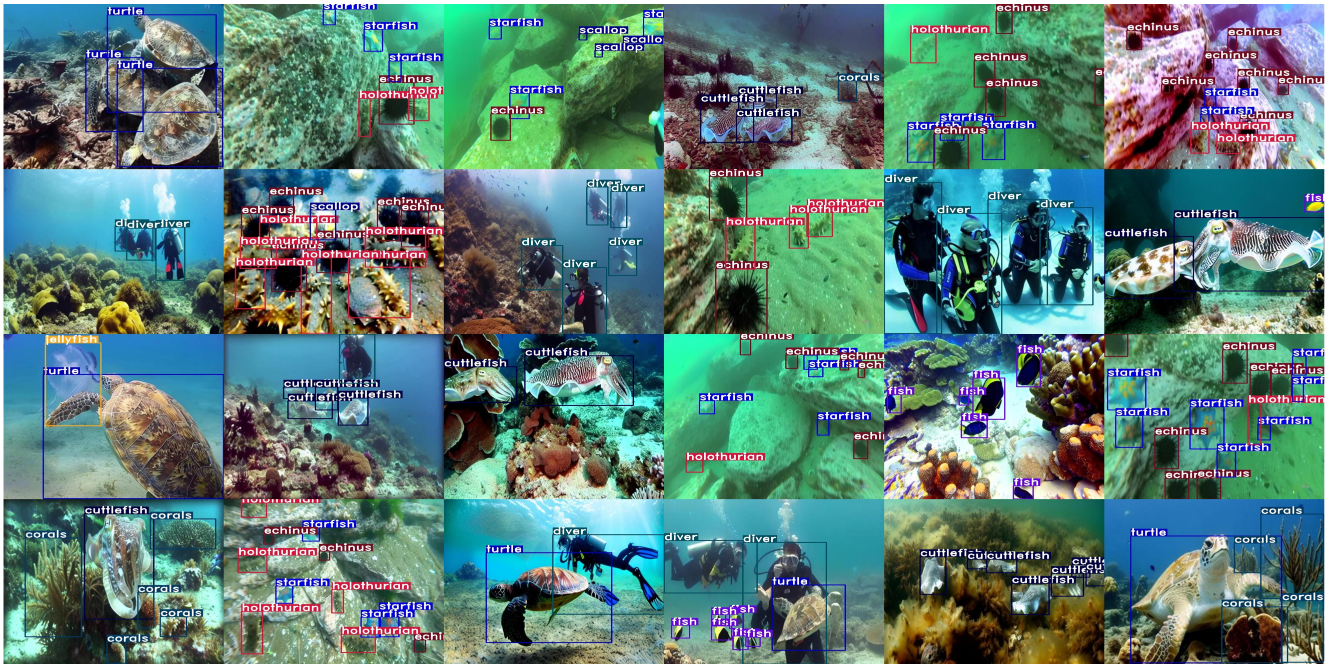

Figure 10. Visualization results generated by our FICGen under the underwater scene (RUOD [6]), with geometric layouts and corresponding object categories superimposed on the generated images.

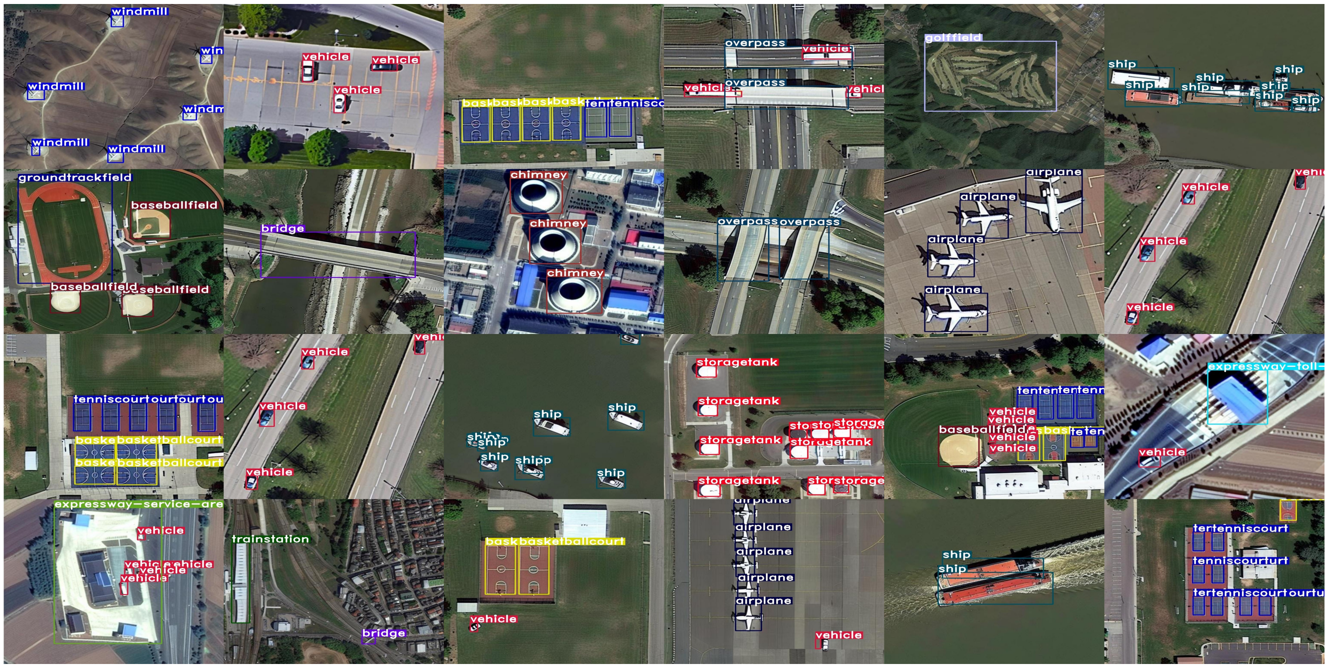

Figure 11. Visualization results generated by our FICGen under the remote sensing scene (DIOR [8]), with geometric layouts and corresponding object categories superimposed on the generated images.

2024. 1, 2, 4, 5

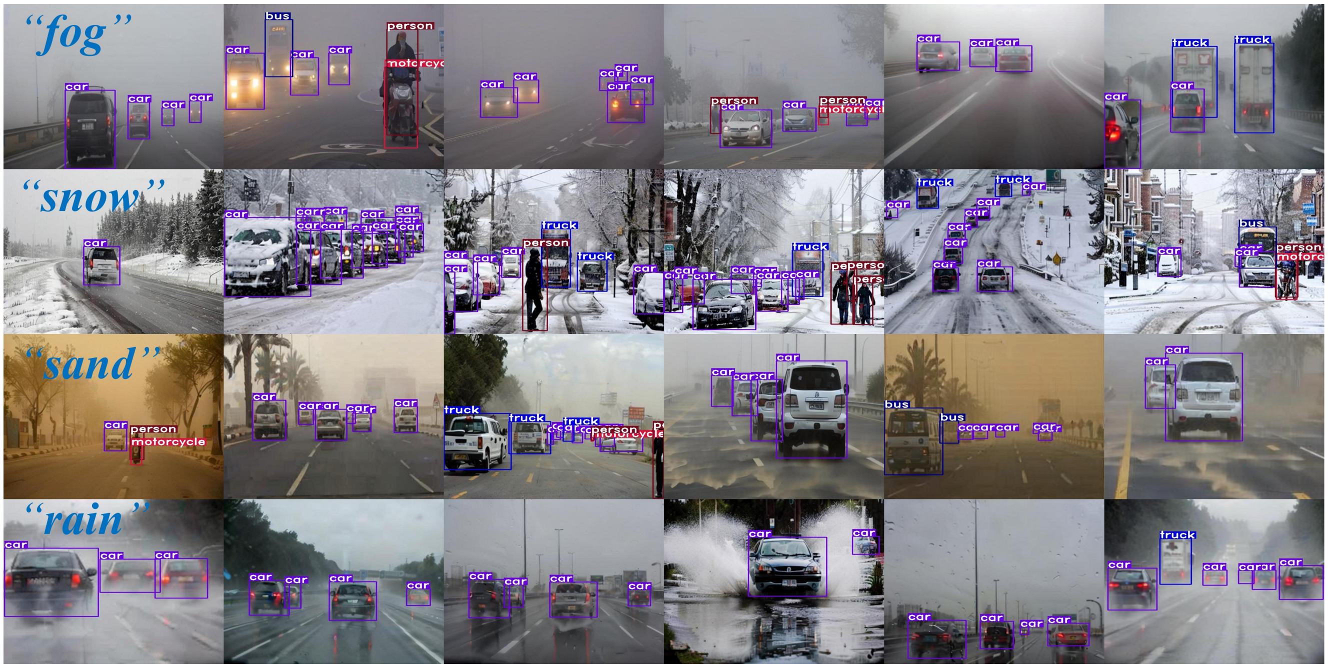

Figure 12. Visualization results generated by our FICGen under the adverse weather condition (DAWN [7]), with geometric layouts and corresponding object categories superimposed on the generated images.