

# Supplementary Materials: Data Generation Scheme for Thermal Modality with Edge-Guided Adversarial Conditional Diffusion Model

Anonymous Authors

## 1 THE SAMPLING PARAMETERS OF DPM-SOLVER++

Table 1 shows the hyperparameters of DPM-solver++ [5] used for speeding up sampling our Edge-guided Conditional Diffusion Model (ECDM) in the Two-stage Modality Adversarial Training (TMAT) strategy. We utilize DPM-solver++ solely during the training phase.

**Table 1: The parameters in dpm-solver++**

Hyper-parameters	Value
timesteps	5
order	3
skip type	time uniform
sampling method	adaptive
type	taylor
condition scale	0.5
absolute tolerance	0.0078
relative tolerance	0.05

## 2 EXPLORATION OF THERMAL IMAGE GENERATION USING DIFFERENT CONDITIONS ON DIFFUSION MODEL

We conduct an evaluation of the quality of generated images under various conditions. Our report includes metrics for the quality of generated images under the following conditions: no condition (marked as NO), thermal image condition (marked as  $\mathcal{D}_{llvip}^{tir}$ ), thermal edge condition (marked as  $\zeta_{llvip}^{tir}$ ), visible image in nighttime (marked as  $\mathcal{D}_{llvip}^{vis}$ ), visible edge image in nighttime (marked as  $\zeta_{llvip}^{vis}$ ), visible image in daytime (marked as  $\mathcal{D}_{prw}$ ), and visible edge images in daytime (marked as  $\zeta_{prw}$ ).  $\mathcal{D}_{llvip}^{vis}$  and  $\zeta_{llvip}^{tir}$  have the same distribution with the thermal domain.  $\mathcal{D}_{llvip}^{vis}$  and  $\zeta_{llvip}^{vis}$  is strictly aligned in time and space with  $\mathcal{D}_{llvip}^{vis}$  and  $\zeta_{llvip}^{tir}$ , so it has different distribution with thermal domain but has same semantic information.  $\mathcal{D}_{prw}$  and  $\zeta_{prw}$  neither has same distribution nor semantic information.

In this experiment, we maintain the training setting identical to the sampling condition. To ensure a fair comparison, we set  $S_{diff} = 70$ . The results are presented in Table 2 and visualized in Figure 1. When no condition is applied to control the generated content, the generated images exhibit a high FID-C score of 250.82 and lack meaningful content. By incorporating conditions, we observe a significant reduction in the FID-C score and improved control over the generated image content. Thermal domain conditions outperform visible domain conditions due to their similar distribution

NO



$\mathcal{D}_{llvip}^{tir}$



$\zeta_{llvip}^{tir}$



$\mathcal{D}_{llvip}^{vis}$



$\zeta_{llvip}^{vis}$



**Figure 1: Visualization of images generated under different conditions.**

**Table 2: Ablation study for different conditions**

Condition	FID↓	FID-C↓	FID-C <sub>clip</sub> ↓	KID↓
NO	257.14	250.82	38.59	0.2817
$\mathcal{D}_{llvip}^{tir}$	67.64	62.53	15.17	0.0408
$\zeta_{llvip}^{tir}$	35.07	35.69	16.15	0.0193
$\mathcal{D}_{llvip}^{vis}$	130.91	133.53	26.53	0.0967
$\zeta_{llvip}^{vis}$	139.91	147.09	26.98	0.1167

with the target domain. Notably,  $\zeta_{llvip}^{tir}$  performs better than  $\mathcal{D}_{llvip}^{vis}$ , as the texture information in  $\mathcal{D}_{llvip}^{vis}$  adversely affects the fine control of edge information in the generated image boundaries. However,  $\zeta_{llvip}^{vis}$  performs relatively poorly compared to  $\mathcal{D}_{llvip}^{vis}$ , since the visible images in the LLVIP dataset [3] are captured at night, resulting in scarce edge information in these images. This finding verifies the importance of edge information in precisely generating fine-granularity content of objects.

## 3 MORE SHOWCASES OF ECDM ON THERMAL OBJECT DETECTION

We also train Faster RCNN [6] with diverse augmentation multiple ratios and mixed ratios. Figure 4 and Figure 5 show some curves

in the experiments. The results are shown in Figures 2 and 3, respectively. Note that the best augmentation multiple ratio at 0.8 in Faster RCNN is 0.8, which achieves a 2.1 improvement on mAP.

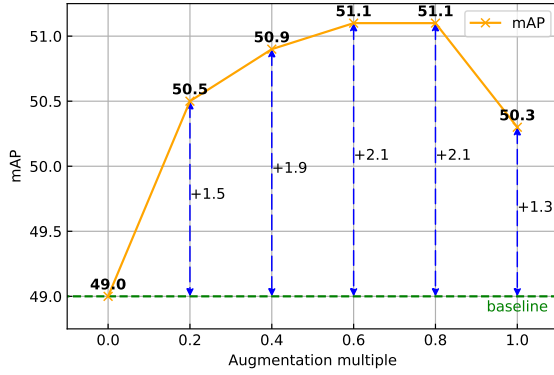


Figure 2: The performance of Faster RCNN trained with various amounts of generated pseudo training data. The x-axis indicates the augmentation multiple. For example, 0.2 indicates that the generated pseudo training data in the entire training sample is only 20% of the real data.

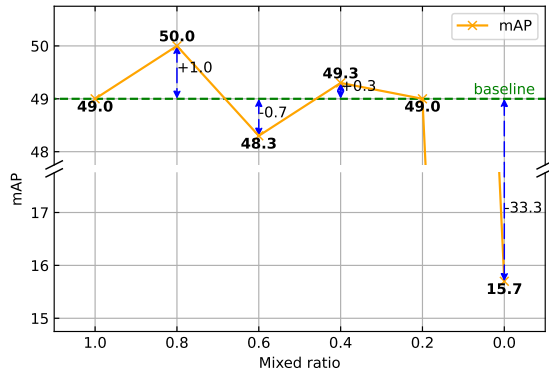


Figure 3: The performance of Faster RCNN trained with various amounts of generated pseudo training data. The x-axis indicates the mixed ratios. For example, 0.2 indicates that the entire training samples have 20% generated pseudo training data and 80% real data.

#### 4 CLASS-WISE RESULTS ON THE FLIR DATASET

We train various object detectors on the FLIR dataset [2], including Faster RCNN [6], RetinaNet [4], CenterNet [1], VNet [8], and DINO [7]. For a fair comparison, we maintain an augmentation multiple ratio of 1.0 throughout this experiment. The FLIR dataset encompasses 15 categories, but we only utilize 5 categories in our

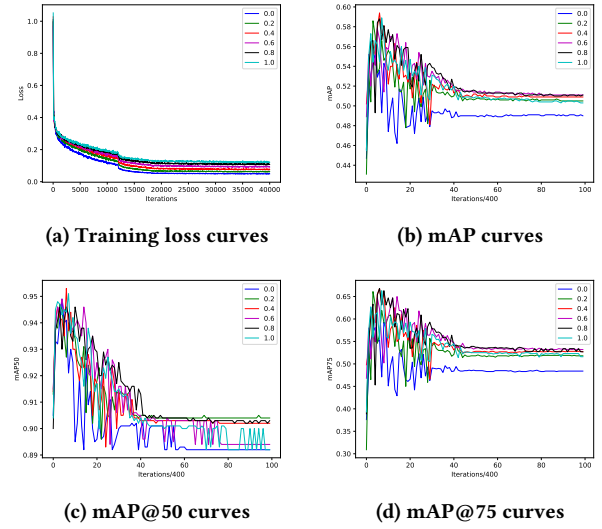


Figure 4: Some curves at different augmentation multiple ratios.

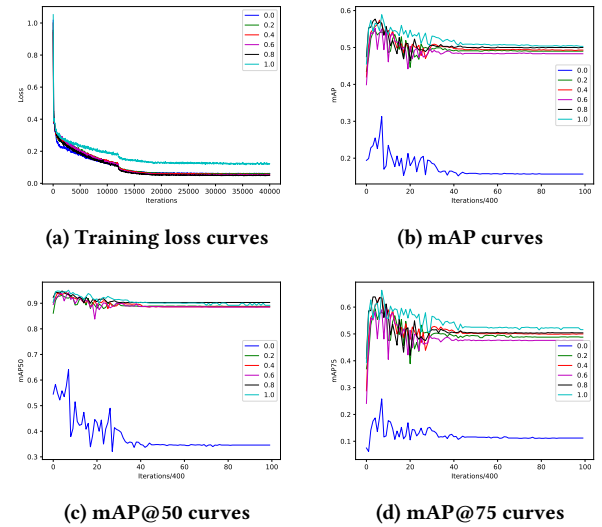


Figure 5: Some curves at different mixed ratios.

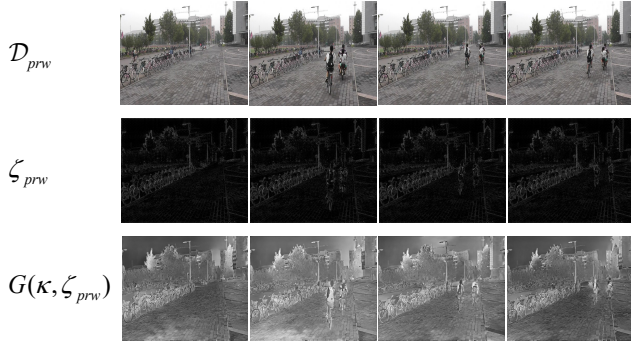
experiments due to the limited labeling. The primary metrics of mAP are presented in the manuscript. We provide class-wise sub metrics of mAP in Table 3.

#### 5 MORE QUALITATIVE RESULTS

We provide more qualitative comparison results with other methods in Figure 8.

The generated samples under the PRW dataset are shown in Figure 6. Figure 6 demonstrates that the generated thermal images exhibit similar overall gray distributions in the global space. However, some discrepancies are observed in specific details, such as the

heads or legs of humans, and bags. These differences highlight the difficulty of the transferability models challenge, owing to the substantial gap in data distribution when generating infrared images from edge images sourced from different datasets.



**Figure 6: Here are some examples of images in the PRW dataset, edge images extracted from images and generated pseudo thermal images under edge images.**

Some failed cases are shown in Figure 7.

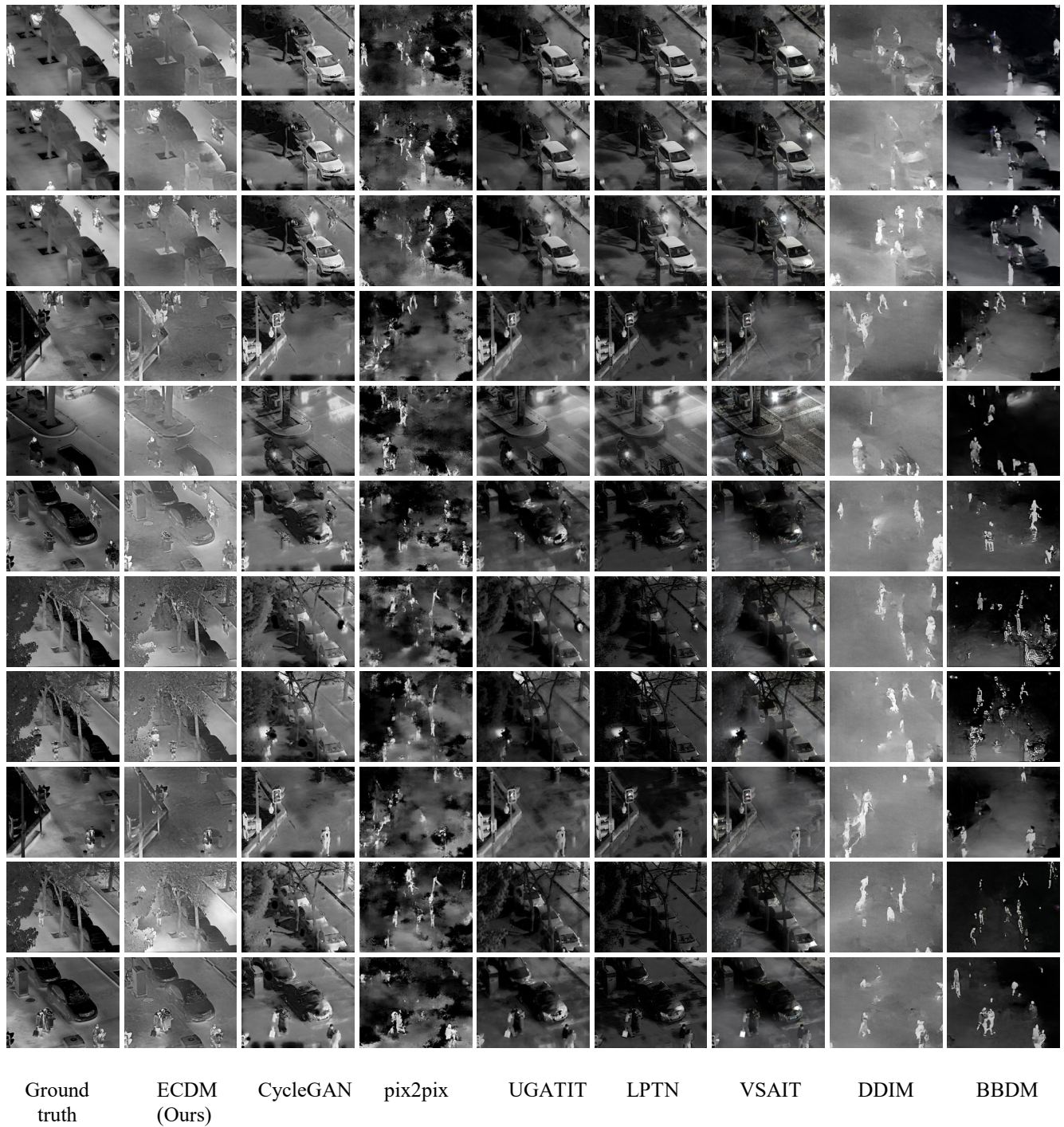


**Figure 7: Typical FAKE thermal images. (a) Blur ghost, which means exits some blurry artifacts in the images. (b) Error color levels, which means images have incorrect color levels. (c) Polarity reversal, which means a hot object has a lower gray value than a cool object (face and cloth).**

## REFERENCES

- [1] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. 2019. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6569–6578.
- [2] Teledyne FLIR. 2019. Free Teledyne FLIR thermal dataset for algorithm training. <https://www.flir.com/oem/adas/adas-dataset-form>. Accessed:2023-08-01.
- [3] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. 2021. LLVIP: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3496–3504.
- [4] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.
- [5] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2023. DPM-Solver++: Fast Solver for Guided Sampling of Diffusion Probabilistic Models. arXiv:2211.01095 [cs.LG]
- [6] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015).
- [7] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. 2022. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. arXiv:2203.03605 [cs.CV]
- [8] Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sunderhauf. 2021. Varifocal-net: An iou-aware dense object detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8514–8523.





**Figure 8: More qualitative comparison of our proposed method with other state-of-the-art methods on the LLVIP test dataset. To ensure fairness and randomness, we use Python’s random module with a fixed seed (1234) to select images from the dataset. The selected images are ‘190065.jpg’, ‘190072.jpg’, ‘190127.jpg’, ‘200143.jpg’, ‘210307.jpg’, ‘230422.jpg’, ‘240321.jpg’, ‘240409.jpg’, ‘260211.jpg’, ‘260304.jpg’, ‘260379.jpg’.**

Table 3: Class-wise mAP results on the FLIR dataset

Method	Class	Pseudo data	mAP	AP@50	AP@75	AP <sub>s</sub>	AP <sub>m</sub>	AP <sub>l</sub>
Faster RCNN	Person	✗	26.1	50.4	24.5	19.1	53.9	54.3
		✓	26.8 (+0.7)	51.7 (+1.3)	25.1 (+0.6)	20.1 (+1.0)	54.5 (+0.6)	57.5 (+3.2)
	Bike	✗	22.3	43.6	20.2	12.1	31.7	35.3
		✓	25.7 (+3.4)	45.1 (+1.5)	28.0 (+7.8)	11.8 (-0.3)	37.8 (+6.1)	25.2 (-0.1)
	Car	✗	43.1	66.9	46.6	23.7	66.8	83.6
		✓	46.0 (+2.9)	69.9	50.1 (+3.5)	26.6 (+2.9)	68.9 (+2.1)	84.5 (+0.9)
	Light	✗	10.0	25.6	4.9	9.4	31.4	-
		✓	10.7 (+0.7)	26.8 (+1.2)	6.5 (+1.6)	10.0 (+0.6)	33.9 (+2.5)	-
	Sign	✗	14.6	24.7	16.3	12.4	42.8	-
		✓	16.6 (+2.0)	27.8 (+3.1)	17.9 (+1.6)	13.8 (+1.4)	50.3 (+7.5)	-
RetinaNet	Person	✗	14.8	37.8	8.9	7.6	40.5	46.8
		✓	16.1 (+1.3)	39.6 (+1.8)	10.9 (+2.0)	8.2 (+0.6)	44.0 (+3.5)	52.6 (+5.8)
	Bike	✗	14.5	33.7	10.1	5.5	23.0	40.4
		✓	15.9 (+1.4)	36.1 (+2.4)	11.7 (+1.6)	4.3 (-1.2)	26.2 (+3.2)	45.4 (+5.0)
	Car	✗	35.5	57.9	36.6	9.9	64.4	81.7
		✓	35.5	58.8 (+0.9)	36.7 (+0.1)	10.2 (+0.3)	64.4	81.3 (-0.4)
	Light	✗	2.6	7.2	1.3	1.5	24.8	-
		✓	2.3 (-0.3)	6.9 (-0.3)	1.2 (-0.1)	1.6 (+0.1)	24.6 (-0.2)	-
	Sign	✗	5.1	10.4	4.9	2.2	40.3	-
		✓	5.9 (+0.8)	12.6 (+2.2)	5.4 (+0.5)	2.9 (+0.7)	42.4 (+2.1)	-
CenterNet	Person	✗	26.2	57.5	20.8	20.4	53.1	53.7
		✓	28.9 (+2.7)	60.3 (+2.8)	24.1 (+3.3)	22.2 (+1.8)	54.7 (+1.6)	59.5 (+5.8)
	Bike	✗	22.6	39.9	22.7	7.3	36.2	32.3
		✓	25.5 (+2.9)	45.0 (+5.1)	23.7 (+1.0)	11.1 (+3.8)	37.5 (+1.3)	35.3 (+3.0)
	Car	✗	45.5	72.7	46.5	23.7	69.6	85.5
		✓	47.6 (+2.1)	74.6 (+1.9)	49.2 (+2.7)	26.3 (+2.6)	70.9 (+1.3)	85.8 (+0.3)
	Light	✗	14.7	39.1	6.6	14.1	38.1	-
		✓	15.7 (+1.0)	42.8 (+2.7)	7.7 (+1.1)	15.2 (+1.1)	37.2 (-0.9)	-
	Sign	✗	18.3	35.6	17.3	15.9	48.5	-
		✓	19.2 (+0.9)	38.0 (+2.4)	17.5 (+0.2)	16.6 (+0.7)	51.4 (+2.9)	-
VFNet	Person	✗	16.0	40.2	10.2	10.3	40.0	40.6
		✓	15.0 (-1.0)	37.9 (-2.3)	9.9 (-0.3)	8.7 (-1.6)	41.0 (+1.0)	44.4 (+3.8)
	Bike	✗	12.2	30.3	6.4	2.9	21.2	3.6
		✓	10.2 (-2.0)	25.6 (-4.7)	5.7 (-0.7)	2.6 (-0.3)	17.6 (-3.6)	15.1 (+11.5)
	Car	✗	35.7	61.1	36.8	15.7	60.1	75.1
		✓	32.5 (-3.2)	57.1 (-4.0)	33.3 (-3.5)	13.2 (-2.5)	56.1 (-4.0)	70.7 (-4.4)
	Light	✗	4.5	12.8	2.0	3.7	27.7	-
		✓	2.8 (-1.7)	8.0 (-4.0)	1.6 (-0.4)	2.3 (-1.4)	19.3 (-8.4)	-
	Sign	✗	7.2	15.8	5.6	4.9	36.4	-
		✓	4.4 (-2.8)	10.3 (-5.5)	3.1 (-2.5)	2.4 (-2.5)	29.7 (-6.7)	-
DINO	Person	✗	12.4	28.2	9.0	11.0	18.5	21.2
		✓	16.5 (+4.1)	39.3 (+11.1)	10.8 (+1.8)	13.4 (+2.4)	29.5 (+11.0)	32.2 (+11.0)
	Bike	✗	3.3	6.9	2.8	0.2	7.6	2.0
		✓	3.5 (+0.2)	8.3 (+1.4)	1.6 (-1.2)	0.7 (+0.5)	6.3 (+1.3)	0.7 (-1.3)
	Car	✗	17.8	33.4	17.2	11.2	28.1	27.4
		✓	20.2 (+2.4)	40.3 (+6.9)	18.0 (+0.8)	12.5 (+1.3)	32.2 (+4.1)	37.3 (+9.9)
	Light	✗	2.8	7.0	1.7	2.7	8.5	-
		✓	3.6 (+0.8)	9.2 (+2.2)	2.1 (+0.4)	3.5 (+1.3)	7.6 (-0.9)	-