Taming Generative Synthetic Data for X-ray Prohibited Item Detection

Jialong Sun, Hongguang Zhu, Weizhe Liu, Yunda Sun, Renshuai Tao and Yunchao Wei

***Abstract*—Training prohibited item detection models requires a large amount of X-ray security images, but collecting and annotating these images is time-consuming and laborious. To address data insufficiency, X-ray security image synthesis methods composite images to scale up datasets. However, previous methods primarily follow a two-stage pipeline, where they implement labor-intensive foreground extraction in the first stage and then composite images in the second stage. Such a pipeline introduces inevitable extra labor cost and is not efficient. In this paper, we propose a one-stage X-ray security image synthesis pipeline (Xsyn) based on text-to-image generation, which incorporates two effective strategies to improve the usability of synthetic images. The Cross-Attention Refinement (CAR) strategy leverages the cross-attention map from the diffusion model to refine the bounding box annotation. The Background Occlusion Modeling (BOM) strategy explicitly models background occlusion in the latent space to enhance imaging complexity. To the best of our knowledge, compared with previous methods, Xsyn is the first to achieve high-quality X-ray security image synthesis without extra labor cost. Experiments demonstrate that our method outperforms all previous methods with 1.2% mAP improvement, and the synthetic images generated by our method are beneficial to improve prohibited item detection performance across various X-ray security datasets and detectors.**

***Index Terms*—Image Generation, Synthetic Data, X-ray Security Image Synthesis, X-ray Prohibited Item Detection.**

I. Introduction

AUTOMATIC prohibited item detection [1]–[7] aims to detect all contraband from a single X-ray security image. Training such models usually requires a large amount of annotated data, but both collecting and annotating X-ray security images are time-consuming and laborious, resulting in a high labor cost for obtaining well-annotated images. For example, collecting a single image from X-ray security inspection equipment can take up to one minute, and multiple rounds of iterative labeling by professional annotators further increase time costs.

To reduce the cost of collecting hand-annotated X-ray security images, utilizing synthetic data has emerged as an effective way. Previous X-ray image synthesis methods mainly utilize two methods to synthesize images: Threat Image Projection-based (TIP-based) synthesis [8], [9] and Generative Adversarial Network-based (GAN-based) synthesis [10]–[14]. 1) TIP-based synthesis involves fusing the prohibited item with the background image through morphological operations [8] or an image fusion neural network [9]. However, it either requires laborious mask annotation for foreground extraction or time-consuming Foreground Threat Image (FTI) collection for fusion network training. 2) GAN-based synthesis enriches the foreground diversity by adopting GAN [10] to generate prohibited items with varying poses and shapes. However, training GAN on foreground images also brings inevitable extra labor cost on data collection and annotation (*e.g*, FTI collection [11], trimap [13], and semantic label [14]).

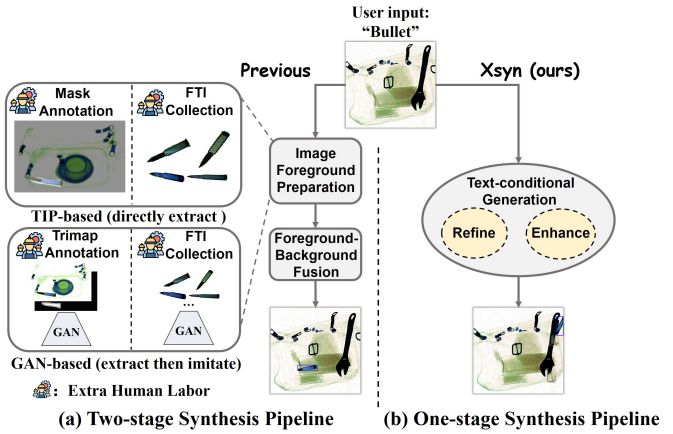

Fig. 1: Analysis of existing X-ray security image synthesis methods. Previous two-stage synthesis methods introduce inevitable labor cost in the first stage (*e.g*, foreground preparation process), which hinders the efficiency of the whole synthesis pipeline. In contrast, Xsyn is a simple and effective one-stage synthesis pipeline, which can automatically refine the synthetic annotation and enhance the synthetic complexity, thereby generating high-quality synthetic data and eliminating extra labor costs.

As shown in Figure 1, we analyze existing X-ray security image synthesis methods, observing that there is one common limitation in previous methods: *they all suffer from inevitable extra labor (*e.g*, FTI collection and annotation)*. We argue that this limitation stems from the fact that previous methods primarily follow a two-stage synthesis pipeline, where the first stage involves extracting foregrounds for the second synthesis stage, thus introducing inevitable extra labor shown in Figure 1 (a). For instance, TIP-based methods directly extract image foregrounds, and GAN-based methods imitate these foregrounds on the basis of extraction. Therefore, the question arises: *Can we achieve high-quality X-ray security image synthesis without extra labor?*

In this paper, we propose a simple and effective one-stage

Jialong Sun, Weizhe Liu, Renshuai Tao and Yunchao Wei are with Institute of Information Science, Beijing Jiaotong University, Beijing, 100044, China (email: {sunjialong, liuweizhe, rstao, yunchao.wei}@bjtu.edu.cn).

Hongguang Zhu is with Faculty of Data Science, City University of Macau, China. (email: zhuhongguang1103@gmail.com).

Yunda Sun is with Nuctech Company Limited, Beijing, 100083, China (email: sunyunda@nuctech.com).

X-ray security image **syn**thesis (**Xsyn**) pipeline to eliminate extra labor cost. The basic idea is illustrated in Figure 1 (b). Our method is based on the text-grounded inpainting pipeline, which requires no extra labor cost and can generate high-quality X-ray security images by bridging the generative power of the diffusion model and the perception capability of SAM [15]. Specifically, we fine-tune the layout-to-image diffusion model through text-grounded inpainting training and then inpaint X-ray security images by providing grounding conditions (*e.g*, bounding boxes with class names). To refine synthetic annotations of the generated X-ray security images, we propose **C**ross-**A**ttention **R**efinement (**CAR**), which refines the bounding box through the cross-attention map from the diffusion model. By designing a median point sampling strategy based on the most class-discriminative part of the cross-attention map, we augment the bounding box prompt and input it to SAM, thus obtaining precise position prediction. Considering the common background occlusion in real-world baggages, we further introduce **B**ackground **O**cclusion **M**odeling (**BOM**) to enhance synthetic complexity, which explicitly models background occlusion in the latent space of the diffusion model. We propose to automatically search the background occluder and then fuse the background occluder with the foreground parts of the latent at the end of the denoising process. With the above strategies, our synthesis method can generate high-quality X-ray security images without labor-intensive cost. These synthetic images can be used to train prohibited item detection models, supplementing real images. To summarize, our contributions are threefold:

- We propose Xsyn, a simple and effective one-stage synthesis pipeline in the X-ray security domain. To the best of our knowledge, Xsyn is the first to achieve high-quality X-ray security image synthesis without incurring additional labor-intensive foreground preparation.
- We present two effective strategies to enhance the usability of synthetic data. The CAR strategy automatically refines the synthetic image annotations, and the BOM strategy explicitly models the background occlusion in X-ray security images to enhance their imaging complexity.
- Experiments on public X-ray security datasets demonstrate that the generated images from our synthesis pipeline are beneficial to improve prohibited item detection performance.

II. Related Work

X-ray Security Image Synthesis. Prohibited item detection models require a large amount of data. Considering the training need, X-ray security image synthesis [8], [9], [11]–[14] has emerged as an effective way to deal with data insufficiency. It can mainly be categorized into two ways: TIP-based synthesis [8], [9] and GAN-based synthesis [11]–[14]. 1) TIP-based synthesis augments X-ray imagery datasets by superimposing prohibited items onto available X-ray security baggage images. For example, TIP [8] blends isolated threat objects onto benign X-ray images through multistage morphological operations and composition. RWSC-Fusion [9] trains an end-to-end region-wise style-controlled fusion network that superimposes prohibited items onto normal X-ray security images to synthesize realistic composite images. 2) GAN-based synthesis aims to directly generate prohibited items. Yang [13] proposes to extract prohibited items with KNN-matting [16] and improve CT-GAN [17] for prohibited item generation. Li [14] presents a GAN-based method for synthesizing X-ray security images with multiple prohibited items by establishing a semantic label library. Zhu [11] propose an improved Self-Attention GAN (SAGAN) [18] to generate diverse X-ray images of prohibited items and integrate them with background images. However, the aforementioned methods all suffer from inevitable extra labor costs, including time-consuming FTI collection [8], [9], [11], mask [8], trimap [12], [13], and semantic label [14] annotation cost. In contrast to previous methods, our method removes extra labor costs and can generate high-quality X-ray security images through an automatic synthesis pipeline.

Generative Data Synthesis for Detection. A series of methods [19]–[23] have utilized generative models for detection data generation in the natural image domain, and can mainly be divided into two manners [19]: copy-paste synthesis [20], [22] and layout-to-image (L2I) generation [19], [23], [24]. 1) Copy-paste synthesis aims to generate separate foreground objects and fuse them with background images. Ge [20] decouples detection data generation into foreground object mask generation and background image generation through DALL-E [25]. Zhao [22] leverages CLIP [26] and Stable Diffusion [27] to obtain images with accurate categories for copy-paste synthesis. However, copy-paste synthesis requires separate foreground image generation, which can bring inevitable extra labor costs in the X-ray security domain. 2) The L2I methods, on the other hand, directly generate the whole image with objects from the layout instruction (*e.g*, bounding boxes with object categories), avoiding the need to generate foregrounds separately. To achieve better controllable generation, GLIGEN [24] integrates a novel gated self-attention mechanism into text-to-image diffusion models for better layout control. GeoDiffusion [19] further translates geometric conditions into text prompts to generate high-quality detection data. To eliminate the extra labor cost, our method is built upon layout-to-image generation, but extends it into text-grounded inpainting to deal with the background distribution discrepancy in the X-ray security domain, and distinctively proposes two effective strategies to improve the usability of generated images.

III. Preliminary

Latent Diffusion Model [27] is a kind of diffusion model that performs the diffusion process in the latent space for text-to-image generation. Specifically, given a noisy latent $\mathbf{z}_t \in \mathbb{R}^{H^{'} \times W^{'} \times C}$ at each timestep $t \in \{0, ..., T-1\}$, a denoising UNet [28] $\epsilon_\theta(\cdot)$ is trained to recover its clean version $\mathbf{z}_0$ by predicting the added noise, and the training objective can be formulated as follows:

$$\mathcal{L}_{LDM} = \mathbb{E}_{\mathbf{z}_0, \epsilon \sim \mathcal{N}(0,1), t} \|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{c})\|^2 \tag{1}$$

where ϵ is the added random Gaussian noise and $\mathbf{c}$ is the generalized condition. For text-to-image generation, $\mathbf{c}$ is the text

Fig. 2: Qualitative comparisons between L2I generation and grounded inpainting. The background of the L2I-generated image (middle) differs a lot from the real-world baggage (left), which may hinder the detection performance. Therefore, we choose grounded inpainting (right) to retain the background.

prompt which will be encoded by a pre-trained CLIP [26] text encoder. For layout-to-image generation, $\mathbf{c}$ further incorporates the grounding condition (*e.g*, bounding boxes with categories).

Eq. 1 can be further reformulated to support inpainting tasks. Specifically, given an inpainting mask $\mathbf{m}$ and the input image, the input image latent $\mathbf{z}_0^{input}$ can be extracted by a pre-trained Vector Quantized Variational AutoEncoder (VQ-VAE) [29], and its masked version $\mathbf{z}_0^{mask}$ is the multiplication of $\mathbf{z}_0^{input}$ and $\mathbf{m}^{resize}$, where $\mathbf{m}^{resize}$ is obtained by resizing $\mathbf{m}$ to the latent size. Based on [24], the input for UNet is expanded as $\mathbf{z}_t^{inpaint} = Concat(\mathbf{z}_t, \mathbf{z}_0^{mask}, \mathbf{m}^{resize})$, which is fed into Eq. 1 to replace $\mathbf{z}_t$ for inpainting training. Then, at each sampling step t, the noisy latent $\mathbf{z}_t$ is updated as follows before denoising:

$$\mathbf{z}_t = \mathbf{z}_{t+1} * (1 - \mathbf{m}^{resize}) + \mathbf{z}_t^{input} * \mathbf{m}^{resize} \tag{2}$$

where $\mathbf{z}_t^{input}$ is the noisy version of $\mathbf{z}_0^{input}$.

IV. Methodology

A. Generation Pipeline

Previous L2I methods [19], [23] in the natural image domain directly use the generated images as synthetic data. However, we find that such an approach is not feasible in the X-ray security image domain since the background of the generated image is uncontrollable and its distribution deviates significantly from that of the real background, as shown in Figure 2. To avoid the above problem, we base the generation pipeline on text-grounded inpainting.

In general, given an X-ray security image $I \in \mathbb{R}^{H \times W \times 3}$, a text prompt Y, and a grounding condition G, the text-grounded inpainting process can be formulated as a function $I^* = F(I, Y, G)$. The grounding condition $G = \{(e_i, l_i)\}_{i=1}^{M}$, where e_i represents the textual description of the object (*e.g*, class name), and $l_i = [x_{i,1}, y_{i,1}, x_{i,2}, y_{i,2}]$ denotes the i-th grounding box (*i.e*, top-left and bottom-right coordinates). The output is an image with the grounding region being repainted, as specified by the text prompt Y and the grounding condition G.

To generate a new X-ray security image, we design two kinds of grounding conditions G_{mod} and G_{add} based on the image annotation L, where $L = \{(c_i, b_i)\}_{i=1}^{N}$, c_i represents the class name, and b_i represents the i-th annotation box, sharing

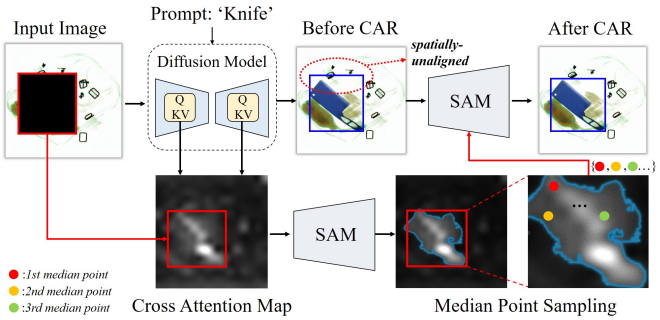

Fig. 3: Cross-Attention Refinement. To obtain the spatial-aligned annotation, we leverage SAM to locate the generated prohibited item based on the rich class-discriminative spatial localization information in the cross-attention map. Please see how the bounding box (blue box) of the generated item is refined.

the same format as the grounding box. Specifically, we first let $G_{mod} = L$ so that we can reuse the annotation and modify the geometry of the original prohibited items (*e.g*, shape and pose). To add a new prohibited item to an image, we first use SAM to segment all elements within the image. Subsequently, we discard the two largest masks by area to prevent out-of-boundary generation. Because they typically correspond to the background and the whole baggage region. Then we select an idle region l_b from the rest masks randomly, and l_b satisfies the following criterion,

$$l_b \in \{l \in S \mid dis(l, b_i) < d, i = 1, 2, \ldots, N\}, \tag{3}$$

where $S = \{s_k\}_{k=1}^{K}$, s_k is the bounding box of the kth object segmented by SAM in image I, $dis(\cdot, \cdot)$ measures the IoU between two bounding boxes and d is the pre-defined threshold. In practice, boxes that are too small will be filtered out. Then we select a category c_b for l_b from a class group which corresponds to specific region areas(refer to the supplementary material) and let $e_b = c_b$ to obtain $G_{add} = \{(e_b, l_b)\}$. By concatenating the class names as the text prompt, we get $Y_{mod} = Concat(\{c_i\}_{i=1}^{N})$ and $Y_{add} = \{e_b\}$. Finally, we can generate a new image in two different ways as follows,

$$\begin{aligned} I^*_{mod} &= F(I, Y_{mod}, G_{mod}), \\ I^*_{add} &= F(I, Y_{add}, G_{add}) \end{aligned} \tag{4}$$

Therefore, we can construct two variants of synthetic data using Eq. 4, named Xsyn-M and Xsyn-A, respectively. This generation pipeline has two advantages. First, it does not require any extra labor cost (*e.g*, FTI collection) compared with previous synthesis methods. Second, it focuses on generating foreground items by altering only a portion of the background, without affecting the overall distribution.

B. Cross-Attention Refinement

Because it is hard for the generated item to be tightly within the grounding box, directly using the grounding box as the annotation box to train detection models will lead to performance degradation of downstream tasks. Instead of

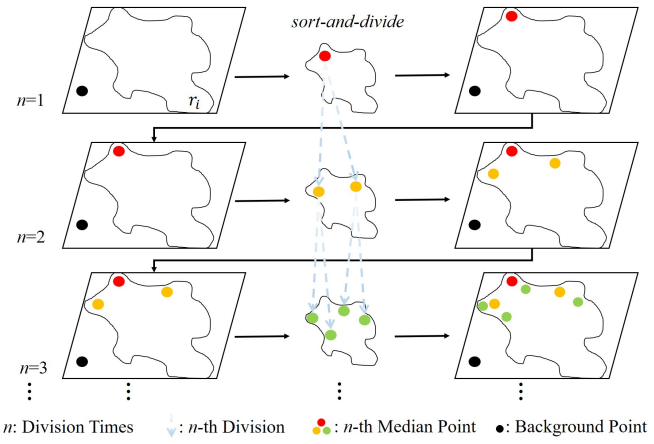

Fig. 4: Median Point Sampling. Considering the background in the bounding box may interfere with the refinement, we propose to enhance the localization by sampling median points as foreground points in a recursive manner. Different colors refer to different division levels.

forcing the generative model to generate spatially aligned items, we retain the generated item and propose CAR to refine its location to obtain the aligned annotation.

Given an input X-ray security image, we first inpaint it using the proposed generation pipeline. Directly using SAM to refine the location by taking the grounding box as input is suboptimal (refer to Table V), because the background can affect the performance of SAM. To address the above issue, we step out of the image domain and propose CAR based on the cross-attention map in the diffusion model. Figure 3 shows the process of CAR. For simplicity, we only discuss the refinement process for one generated item. For the generated item corresponding to $g_i = (e_i, l_i) \in G$, we obtain the average cross-attention map $M_i \in \mathbb{R}^{H \times W}$ from the diffusion model for the text token corresponding to e_i. The CAR process takes as input the generated image I^*, the cross-attention map M_i, and the grounding location l_i. The output is the refined annotation location for the generated item. Specifically, we first obtain the most class-discriminative region r_i by using SAM to segment M_i within l_i. To help SAM better locate the generated item based on l_i, we then propose a median point sampling strategy to sample points P_i from l_i and combine these points with l_i as prompt $\mathcal{P}_i$ for SAM to locate the generated item, where $\mathcal{P}_i = P_i \cup \{l_i\}$.

Median Point Sampling (MPS). Figure 4 depicts the basic idea of median point sampling. We aim to sample foreground points inside r_i and background points outside r_i. Specifically, we choose the point with the minimum activation value outside r_i as the background point p_i^b. To sample foreground points, we first sort all points within r_i by their activation values and choose the median point $p_i^{f_1}$ as the first foreground point. Then we divide r_i into two sub-regions r_i^1 and r_i^2, where the activation values in r_i^1 are all below that of $p_i^{f_1}$, and the activation values in r_i^2 are all above that of $p_i^{f_1}$. By extension, we perform the same *sort-and-divide* operation on the subsequent sub-regions recursively and gather all the median points as foreground points. Therefore, we obtain the final point set $P_i = \{p_i^{f_1}, p_i^{f_2}, \ldots, p_i^{f_{2^n-1}}, \ldots, p_i^{f_{2^n-1}}, p_i^b\}$ which has $2^n - 1$ foreground points and one background point in total, where n indicates the division times. For example, the red, orange, and green points in Figure 4 are in the 1st, 2nd, and 3rd divisions, respectively. We argue that median points describe the central tendency of data points belonging to the prohibited item, which are less affected by extreme activation values in the cross-attention map.

Finally, the refinement process uses SAM to segment I^* by taking $\mathcal{P}_i$ as visual prompts and assigns the bounding box of the segmented region to be the annotation box, thus obtaining more precise location prediction for the generated item. The CAR strategy takes advantage of the segmentation capability of SAM and the cross-attention map of the diffusion model to obtain the refined bounding box annotation. Despite its simplicity, our CAR strategy can achieve automatic annotation refinement that benefits prohibited item detection performance.

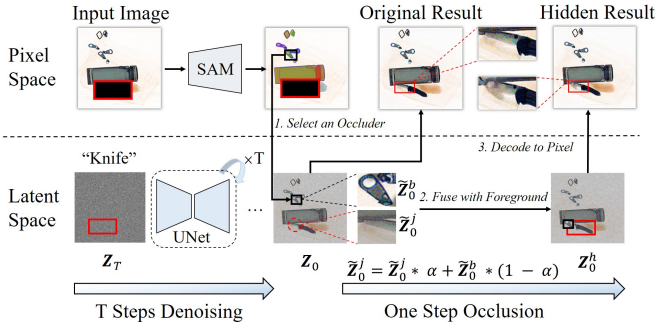

Fig. 5: Background Occlusion Modeling. BOM performs occlusion through regional recombination in the latent space. For simplicity, we omit other variables and components of the diffusion model since the whole generation process has been elaborated.

C. Background Occlusion Modeling

The generated prohibited items are too clear, which is inconsistent with complex real-world occlusion scenarios and may induce overfitting problems for detection. To address the above problem and further enrich the imaging complexity of synthetic images, we simulate the common background occlusion in real baggage by applying background occlusion modeling shown in Figure 5, which fuses the specified background region with foreground regions in the latent space to occlude prohibited items.

Specifically, given an input X-ray security image I, we first select an occluder from the background in pixel space by using SAM to segment every object in I, and use Eq. 3 to determine the location of the occluder[1]. Next, we adopt the proposed generation pipeline to inpaint I but slightly modify the latent sampling process. As shown in Figure 5, a noisy latent $\mathbf{z}_T \in \mathbb{R}^{H' \times W' \times C}$ sampled from the standard normal distribution $\mathcal{N}(0, 1)$ is passed to the denoising UNet,

[1]Eq. 3 used here is reformed as $l_o \in \{l \in S \mid dis(l, b_i) < d, dis(l, l_b) < d, i = 1, 2, \ldots, N\}$ if we use G_{add}.

to obtain the denoised latent $\mathbf{z}_0$ after T steps of denoising. If we directly decode $\mathbf{z}_0$ to the pixel space, then we will get the original result with no occlusion. To occlude the prohibited item, we perform a weighted recombination of the occluder region and foreground regions in latent space for one more step as follows:

$$\tilde{\mathbf{z}}_0^j = \tilde{\mathbf{z}}_0^b * \alpha + \tilde{\mathbf{z}}_0^j * (1 - \alpha) \tag{5}$$

$$\tilde{\mathbf{z}}_0^j = Crop(\mathbf{z}_0, l_j^{'}) \tag{6}$$

$$\tilde{\mathbf{z}}_0^b = Crop(\mathbf{z}_0, l_o^{'}) \tag{7}$$

where $\tilde{\mathbf{z}}_0^j$ and $\tilde{\mathbf{z}}_0^b$ is the j-th occluded foreground region and the occluder region of $\mathbf{z}_0$ respectively. α adjusts the degree of occlusion. $Crop(\cdot)$ represents the process of cropping $\mathbf{z}_0$ to the region corresponding to the occluded region $l_j^{'}$ or the occluder region $l_o^{'}$, where $l_o^{'} = [x_{o,1}^{'}, y_{o,1}^{'}, x_{o,2}^{'}, y_{o,2}^{'}]$, and $l_j^{'}$ can be obtained as follows:

$$l_j^{'} \in \{Re(l_j, l_o^{'}) \mid l_j \in G \cup L, j = 1, 2, \ldots, M + N\} \tag{8}$$

where $Re(\cdot)$ first projects l_j to latent space and then perturbs it as follows:

$$\begin{aligned} x_{j,1}^{'} &= Rand(Max(x_{j,1}^{'} - w_o^{'}, 0), x_{j,2}^{'}), \\ y_{j,1}^{'} &= Rand(Max(y_{j,1}^{'} - h_o^{'}, 0), y_{j,2}^{'}), \\ x_{j,2}^{'} &= Min(x_{j,1}^{'} + w_o^{'}, W^{'}), \\ y_{j,2}^{'} &= Min(y_{j,1}^{'} + h_o^{'}, H^{'}) \end{aligned} \tag{9}$$

where $Rand(\cdot)$ randomly samples an integer between the lower bound and the upper bound. $w_o^{'}$ and $h_o^{'}$ is the width and height of $l_o^{'}$ respectively. We let $l_j^{'} = [x_{j,1}^{'}, y_{j,1}^{'}, x_{j,2}^{'}, y_{j,2}^{'}]$ be the j-th occluded region. The hidden version of $\mathbf{z}_0$ is termed as $\mathbf{z}_0^h$. Finally, we decode $\mathbf{z}_0^h$ to pixel space and obtain the hidden result shown in Figure 5.

Through the regional recombination enabled by BOM, the foreground region can be occluded by a random item from the background, which enhances the imaging complexity of synthetic images. It is worth noting that the original result in Figure 5 is used by CAR to obtain the refined annotation, and we adopt the hidden result as the final synthetic image.

V. Experiments

A. Experimental Setups

Datasets. We conduct experiments on three widely used X-ray security datasets: PIDray [30], OPIXray [31], and HiXray [32]. The details of datasets can be found in the supplementary material.

Implementation Details. *Generation.* We base the generation pipeline on GLIGEN [24]. Specifically, we finetune GLIGEN for 180K steps for text grounded generation training and 50K steps for inpainting training with the batch size set to 8. During inference, we sample images using the DDIM [33] scheduler for 50 steps with the classifier-free guidance (CFG) set as 7.5. *Synthetic images for training.* Taking data annotations of the training set as input, we generate synthetic images using the proposed generation pipeline and apply CAR and BOM to these images. Specifically, we construct two variants of synthetic images, named Xsyn-M and Xsyn-A, respectively. To prevent the disparities in generated data from affecting the model's generalization on real data, we combine the generated images with real images as the final training set, as adopted in DetDiffusion [23]. The spatial resolution of synthetic images is 512×512. *Detection.* We use MMDetection [34] to train downstream detectors. DINO [35] detector with ResNet-50 backbone is used to evaluate the dataset following the default DINO configuration of MMDetection. For all detectors, we uniformly train them for 6 epochs. 4 NVIDIA RTX 3090 GPUs are used for all experiments. More implementation details can be found in the supplementary material.

Evaluation Metrics. Mean average precision (mAP), as the common metric in object detection tasks [36], is used to evaluate the performance. We also evaluate AP for each category and for different occlusion levels on PIDray.

B. Main Results

In this section, we evaluate the performance of the proposed synthesis method for object detection training by supplementing real images with synthetic images generated by our method. To this end, we first compare our approach with previous methods on the PIDray dataset, and then investigate the potential of synthetic data by varying the amount of real images. Finally, we test the effectiveness of our method across various X-ray security datasets and detectors.

Setup. For both Xsyn-M and Xsyn-A generations, we filter out the bounding boxes smaller than a threshold ratio of the image area, and the threshold ratio is 0.1%, 0.4%, and 0.5% for PIDray, OPIXray, and HiXray, respectively. For comparison experiments, we compare Xsyn-M with synthesis methods of the natural image domain and Xsyn-A with previous labor-intensive X-ray security image synthesis methods.

Comparisons with previous methods. Table I shows the results of object detection on PIDray dataset. Our Xsyn-M achieves superior performance compared with methods in the natural image domain, revealing the advantages of the proposed synthesis pipeline. Besides, Xsyn-M can achieve a competitive performance, *i.e*, 69.1% $v.s.$ 69.5% for mAP compared with SAGAN [18], and Xsyn-A can further surpass it by 1.2% mAP. It is worth noting that our synthetic data does not require additional labor compared with previous methods, while data produced by TIP [8], CT-GAN [17], and SAGAN [18] rely on laborious pixel-wise foreground extraction. Both Xsyn-M and Xsyn-A show consistent improvement for almost all classes, especially for some difficult classes (*e.g*, **+7.2%** for Gun with Xsyn-A).

Potential of synthetic data. As shown in Figure 6, we plot the validation mAP curve on PIDray, and the synthetic data generated by our method has better training efficiency compared with previous methods. It indicates that our synthetic data has learned the distribution of X-ray prohibited items and can lead a faster training convergence.

Performance on more datasets and detectors. We extend the evaluation of our method on the OPIXray and HiXray datasets, respectively. The results in Table II demonstrate that our method improves detection performance across various

TABLE I: Comparisons on PIDray dataset. We compare our approach with previous synthesis methods using DINO with ResNet-50 backbone on the PIDray dataset. 'Easy', 'Hard', and 'Hidden' refer to different levels of detection difficulty. 'BA', 'PL', 'HM', 'PO', 'SC', 'WR', 'GU', 'BU', 'SP', 'HA', 'KN' and 'LI' suggest Baton, Pliers, Hammer, Powerbank, Scissors, Wrench, Gun, Bullet, Sprayer, HandCuffs, Knife and Lighter, respectively. *: represents the original L2I generation setting.

Method	Average Precision↑																
	mAP	AP_{50}	Easy	Hard	Hidden	BA	PL	HM	PO	SC	WR	GU	BU	SP	HA	KN	LI
Real only	68.4	81.7	74.0	69.7	52.1	76.2	86.1	83.9	74.8	72.1	90.6	29.6	62.2	56.2	89.6	38.7	61.0
TIP [8]	69.0	82.0	74.9	70.9	51.1	75.9	86.4	84.0	74.7	74.5	91.4	27.4	63.2	59.2	89.5	43.1	58.8
CT-GAN [17]	69.4	82.4	75.3	71.0	52.1	75.9	86.4	83.7	74.0	73.2	91.8	35.4	62.2	59.3	90.2	39.5	60.8
SAGAN [18]	69.5	82.2	75.0	70.9	53.5	76.2	88.1	85.0	75.2	74.5	91.7	29.6	62.5	61.7	89.8	40.7	59.5
GeoDiffusion [19]	64.6	78.4	71.6	64.6	47.6	72.6	82.2	78.8	73.6	69.8	88.1	25.1	57.5	56.2	86.7	28.0	56.6
GLIGEN* [24]	64.9	78.6	73.1	65.2	45.3	72.0	83.2	76.6	71.4	69.4	88.0	28.0	57.6	55.6	88.4	32.5	56.1
Xsyn-M	69.1	82.1	75.5	70.8	50.7	73.4	86.5	84.2	75.8	72.9	91.0	35.5	63.6	60.2	89.8	36.1	60.0
Xsyn-A	**70.7**	**83.8**	**76.8**	**71.7**	**54.1**	**76.7**	**85.6**	**85.1**	**76.1**	**74.8**	**91.7**	**36.8**	**64.1**	**63.5**	**89.2**	**44.5**	**60.1**

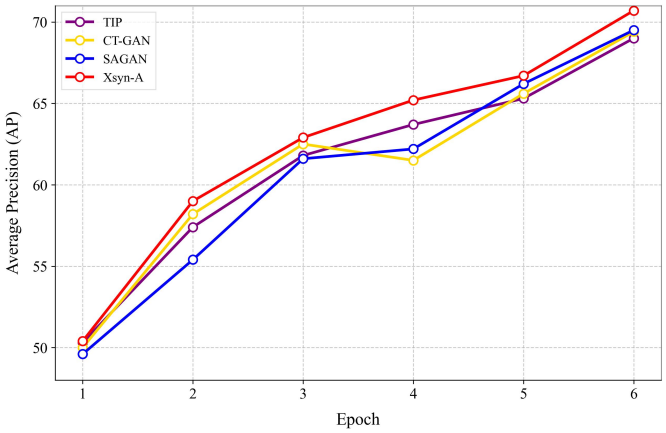

Fig. 6: Potential of synthetic data. Our synthetic data achieves the best detection performance throughout the whole training period.

TABLE II: Performance on OPIXray and HiXray. Our method is effective for various X-ray security datasets.

Dataset	Setting	mAP	AP_{50}	AP_{75}
OPIXray [31]	Real only	39.5	90.2	26.0
	+Xsyn-A	**40.1**	**90.1**	**26.1**
HiXray [32]	Real only	49.3	83.4	53.2
	+Xsyn-A	**50.4**	**83.9**	**55.5**

datasets. We further conduct experiments on various detectors, including CNN-based and Transformer-based [37] architectures, to evaluate the generalization ability. As shown in Table III, our synthetic images achieve consistent improvement regardless of the detection models.

C. Ablation Study

In this section, we conduct ablation studies on the proposed strategies and their specific design choices, respectively. We first ablate the parameter setting of CAR, and then ablate BOM on the basis of CAR. All ablation studies are conducted on the best-performing Xsyn-A, and PIDray dataset is used for all experiments.

TABLE III: Performance on various detectors. Our method can improve prohibited item detection performance consistently, regardless of detectors and backbone architectures.

Type	Stage	Method	Backbone	mAP	AP_{50}	AP_{75}
CNN-based	one	ATSS	R101	65.2	80.8	72.6
		+Xsyn-A	**R101**	**65.5**	**81.3**	**73.0**
	two	C-RNN	R101	68.0	82.6	75.5
		+Xsyn-A	**R101**	**69.1**	**83.4**	**76.4**
		C-RNN	X101	69.6	83.7	77.0
		+Xsyn-A	**X101**	**70.2**	**84.3**	**77.4**
Transformer-based		DINO	R50	68.4	81.7	73.5
		+Xsyn-A	**R50**	**70.7**	**83.8**	**76.7**
		DINO	Swin	76.1	88.6	81.8
		+Xsyn-A	**Swin**	**78.1**	**89.9**	**83.5**

TABLE IV: Ablation studies on proposed strategies. We first add CAR and then BOM to investigate their performance separately. Best results are achieved when both strategies are adopted.

Method	mAP	AP_{50}	AP_{75}
Real only	68.4	81.7	73.9
+Xsyn-A (w/o CAR)	69.6	82.3	75.5
+Xsyn-A (w/o BOM)	70.3	83.1	76.0
+Xsyn-A	**70.7**	**83.8**	**76.7**

The proposed strategies. Table IV presents the impact of the proposed strategies. We analyze the effect of each proposed strategy by sequentially adding 1) CAR and 2) BOM. The results demonstrate the relative importance of each strategy, with all strategies performing the best.

Hyper-parameters of proposed strategies. *Median point sampling.* Table V (upper) shows the performance of CAR using different division times n, where $n = 0$ means that we only use the grounding box to implement refinement. The gain reaches its biggest when $n = 4$, indicating the benefit of incorporating median points and suggesting that MPS has good scalability for annotation refinement. We set n to 4

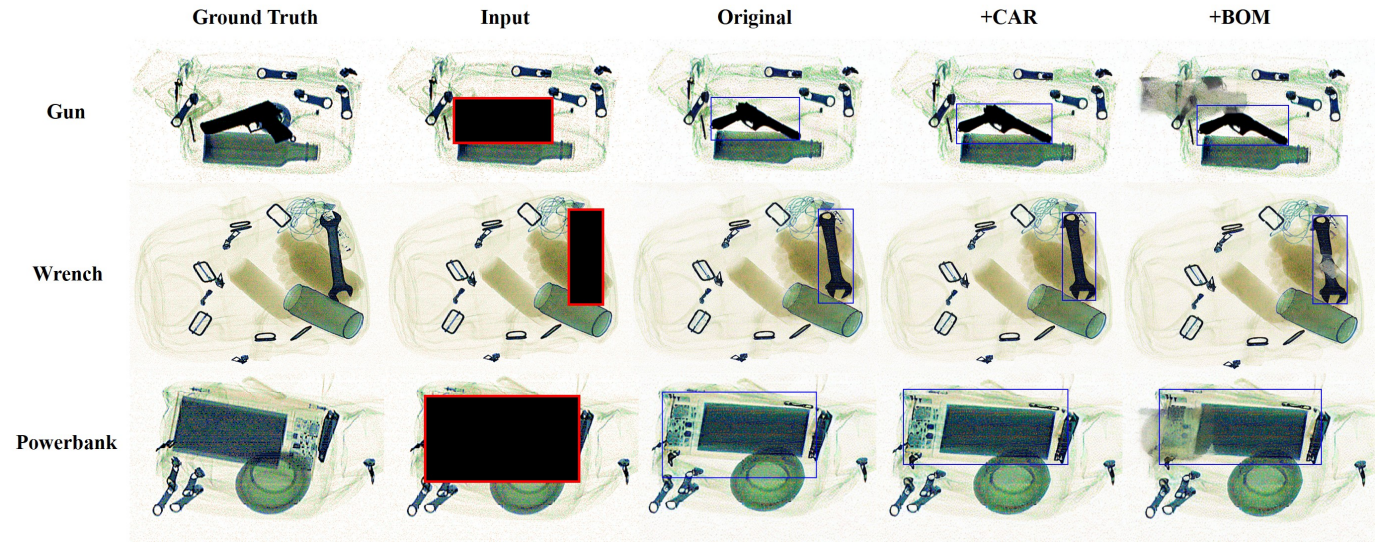

Fig. 7: Qualitative results on PIDray dataset. Our method can synthesize well-annotated and realistic X-ray security images. The blue boxes in the 3rd column and the last two columns refer to the input grounding boxes and the refined annotation boxes, respectively. Please zoom in for better visualization.

TABLE V: Ablations on hyper-parameters of proposed strategies. We ablate the choice of division times *n* for CAR and latent occlusion coefficient α for BOM respectively on Xsyn-A.

Type	Setting	mAP	AP_{50}	AP_{75}
CAR-n	0	69.7	82.5	75.6
	1	69.9	82.7	75.9
	2	70.1	83.0	75.9
	3	70.2	82.8	76.0
	4	**70.3**	**83.1**	**76.0**
BOM-α	0.1	70.3	83.1	76.3
	0.3	**70.7**	**83.8**	**76.7**
	0.5	70.2	82.8	76.0
	0.7	69.8	82.5	75.5

TABLE VI: BOM ablations on occlusion space and period.

	Period	mAP	AP_{50}	AP_{75}
Latent Space	t	69.9	82.9	75.5
	T	**70.7**	**83.8**	**76.7**
Pixel Space	-	69.9	82.7	75.8

for other experiments. *Latent occlusion coefficient.* Table V (bottom) provides the ablation study for occlusion coefficient α. The performance increases when α changes from 0.1 to 0.3, while it decreases from 0.3 to 0.7. The result suggests that a medium occlusion coefficient is beneficial to enhance the imaging complexity, while a too small or too large occlusion coefficient cannot model the complex occlusion in real-world baggage. Therefore, the optimum α is set to 0.3 for better imaging complexity enhancement.

Occlusion space and period. The ablation study for occlusion space and period is shown in Table VI. We fuse the occluder region with foreground regions in the original image to implement occlusion modeling in pixel space. The result shows that modeling occlusion in latent space achieves better performance than in pixel space. We also investigate the influence of the occlusion period by modeling occlusion at each denoising step t, but the performance is much lower than the original version. We argue that such an approach may destroy the distribution of foregrounds in the cross-attention map, thus affecting the process of CAR.

D. Qualitative Results

We provide qualitative results on the PIDray dataset shown in Figure 7. The original results in the 3rd column have obvious spatial misalignment between the generated prohibited item and the bounding box. When we apply CAR to the original results, the bounding box is refined to enclose the prohibited item tightly, as shown in the 4th column. We further enhance the imaging complexity in the 5th column by using BOM to occlude the prohibited item. It is worth mentioning that we perform CAR on the original image and apply BOM to obtain the hidden image, which ensures that the annotation refinement will not be compromised by the introduction of occlusion.

VI. Conclusion

In this paper, we propose Xsyn, a simple and effective one-stage X-ray security image synthesis pipeline to generate high-quality prohibited item detection data. In contrast to the previous two-stage methods, for the first time, our method removes the labor-intensive foreground extraction procedure. To improve the usability of generative synthetic data, our method incorporates two effective strategies to automatically refine the synthetic annotation and enhance the synthetic complexity. The synthetic images generated by our method can improve the prohibited item detection performance across various public datasets and detectors. We hope Xsyn can

bring new inspiration for exploiting the potential of generative synthetic data in the X-ray security domain.

REFERENCES

[1] S. Akcay and T. Breckon, "Towards automatic threat detection: A survey of advances of deep learning within x-ray security imaging," *Pattern Recognition*, vol. 122, February 2022. [Online]. Available: https://breckon.org/toby/publications/papers/akcay22survey.pdf

[2] B. Isaac-Medina, S. Yucer, N. Bhowmik, and T. Breckon, "Seeing through the data: A statistical evaluation of prohibited item detection benchmark datasets for x-ray security screening," in *Proc. Conf. Computer Vision and Pattern Recognition Workshops*. IEEE/CVF, June 2023, pp. 524–533. [Online]. Available: https://breckon.org/toby/publications/papers/isaac23evaluation.pdf

[3] T. Webb, N. Bhowmik, Y. Gaus, and T. Breckon, "Operationalizing convolutional neural network architectures for prohibited object detection in x-ray imagery," in *Proc. Int. Conf. on Machine Learning Applications*. IEEE, December 2021, pp. 610–615. [Online]. Available: https://breckon.org/toby/publications/papers/web21xray.pdf

[4] N. Bhowmik, Y. Gaus, and T. Breckon, "On the impact of using x-ray energy response imagery for object detection via convolutional neural networks," in *Proc. Int. Conf. on Image Processing*. IEEE, September 2021, pp. 1224–1228. [Online]. Available: https://breckon.org/toby/publications/papers/bhowmik21energy.pdf

[5] B. Isaac-Medina, C. Willcocks, and T. Breckon, "Multi-view object detection using epipolar constraints within cluttered x-ray security imagery," in *Proc. Int. Conf. Pattern Recognition*. IEEE, October 2020, pp. 9889–9896. [Online]. Available: https://breckon.org/toby/publications/papers/isaac20multiview.pdf

[6] Y. Gaus, B. Isaac-Medina, N. Bhowmik, Y. Lam, and T. Breckon, "Semi-supervised object-wise anomaly detection for firearm and firearm component detection in x-ray security imagery," in *Proc. Computer Vision Pattern Recognition Workshops*. IEEE/CVF, June 2025, to appear. [Online]. Available: https://breckon.org/toby/publications/papers/gaus25anomaly.pdf

[7] N. Bhowmik and T. Breckon, "Joint sub-component level segmentation and classification for anomaly detection within dual-energy x-ray security imagery," in *Proc. Int. Conf. on Machine Learning Applications*. IEEE, December 2022, pp. 1463–1467. [Online]. Available: https://breckon.org/toby/publications/papers/bhowmik22subcomponent.pdf

[8] N. Bhowmik, Q. Wang, Y. F. A. Gaus, M. Szarek, and T. P. Breckon, "The good, the bad and the ugly: Evaluating convolutional neural networks for prohibited item detection using real and synthetically composited x-ray imagery," *arXiv preprint arXiv:1909.11508*, 2019.

[9] L. Duan, M. Wu, L. Mao, J. Yin, J. Xiong, and X. Li, "Rwsc-fusion: Region-wise style-controlled fusion network for the prohibited x-ray security image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 398–22 407.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[11] Y. Zhu, Y. Zhang, H. Zhang, J. Yang, and Z. Zhao, "Data augmentation of x-ray images in baggage inspection based on generative adversarial networks," *IEEE Access*, vol. 8, pp. 86 536–86 544, 2020.

[12] Z. Zhao, H. Zhang, and J. Yang, "A gan-based image generation method for x-ray security prohibited items," in *Pattern Recognition and Computer Vision: First Chinese Conference, PRCV 2018, Guangzhou, China, November 23-26, 2018, Proceedings, Part I 1*. Springer, 2018, pp. 420–430.

[13] J. Yang, Z. Zhao, H. Zhang, and Y. Shi, "Data augmentation for x-ray prohibited item images using generative adversarial networks," *IEEE Access*, vol. 7, pp. 28 894–28 902, 2019.

[14] D.-s. Li, X.-b. Hu, H.-g. Zhang, and J.-f. Yang, "A gan based method for multiple prohibited items synthesis of x-ray security image," *Opto-electronics Letters*, vol. 17, no. 2, pp. 112–117, 2021.

[15] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[16] Q. Chen, D. Li, and C.-K. Tang, "Knn matting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 9, pp. 2175–2188, 2013.

[17] X. Wei, B. Gong, Z. Liu, W. Lu, and L. Wang, "Improving the improved training of wasserstein gans: A consistency term and its dual effect," *arXiv preprint arXiv:1803.01541*, 2018.

[18] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International conference on machine learning*. PMLR, 2019, pp. 7354–7363.

[19] K. Chen, E. Xie, Z. Chen, Y. Wang, L. Hong, Z. Li, and D.-Y. Yeung, "Geodiffusion: Text-prompted geometric control for object detection data generation," *arXiv preprint arXiv:2306.04607*, 2023.

[20] Y. Ge, J. Xu, B. N. Zhao, N. Joshi, L. Itti, and V. Vineet, "Dall-e for detection: Language-driven compositional image synthesis for object detection," *arXiv preprint arXiv:2206.09592*, 2022.

[21] H. Fang, B. Han, S. Zhang, S. Zhou, C. Hu, and W.-M. Ye, "Data augmentation for object detection via controllable diffusion models," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 1257–1266.

[22] H. Zhao, D. Sheng, J. Bao, D. Chen, D. Chen, F. Wen, L. Yuan, C. Liu, W. Zhou, Q. Chu *et al.*, "X-paste: Revisiting scalable copy-paste for instance segmentation using clip and stablediffusion," in *International Conference on Machine Learning*. PMLR, 2023, pp. 42 098–42 109.

[23] Y. Wang, R. Gao, K. Chen, K. Zhou, Y. Cai, L. Hong, Z. Li, L. Jiang, D.-Y. Yeung, Q. Xu *et al.*, "Detdiffusion: Synergizing generative and perceptive models for enhanced data generation and perception," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7246–7255.

[24] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y. J. Lee, "Gligen: Open-set grounded text-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 511–22 521.

[25] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International conference on machine learning*. Pmlr, 2021, pp. 8821–8831.

[26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[27] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[29] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[30] L. Zhang, L. Jiang, R. Ji, and H. Fan, "Pidray: A large-scale x-ray benchmark for real-world prohibited item detection," *International Journal of Computer Vision*, vol. 131, no. 12, pp. 3170–3192, 2023.

[31] Y. Wei, R. Tao, Z. Wu, Y. Ma, L. Zhang, and X. Liu, "Occluded prohibited items detection: An x-ray security inspection benchmark and de-occlusion attention module," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 138–146.

[32] R. Tao, Y. Wei, X. Jiang, H. Li, H. Qin, J. Wang, Y. Ma, L. Zhang, and X. Liu, "Towards real-world x-ray security inspection: A high-quality benchmark and lateral inhibition module for prohibited items detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 923–10 932.

[33] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2021.

[34] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[35] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," *arXiv preprint arXiv:2203.03605*, 2022.

[36] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[37] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

Jialong Sun is a master student at Beijing Jiaotong University.

Hongguang Zhu is an Associate Professor at City University of Macau.

Weizhe Liu is a master student at Beijing Jiaotong University.

Yunda Sun is a Senior Engineer at Nuctech Company Limited.

Renshuai Tao is an Associate Professor at Beijing Jiaotong University.

Yunchao Wei is a Professor at Beijing Jiaotong University.