Adapting Vehicle Detectors for Aerial Imagery to Unseen Domains with Weak Supervision

Xiao Fang[1], Minhyek Jeon[1], Zheyang Qin[1], Stanislav Panev[1], Celso de Melo[2], Shuowen Hu[2], Shayok Chakraborty[1,3], Fernando De la Torre[1]

[1]Carnegie Mellon University, [2]DEVCOM Army Research Laboratory, [3]Florida State University,

{xfang2, minhyekj, zheyangq, spanev}@andrew.cmu.edu,
{celso.m.demelo.civ, shuowen.hu.civ}@army.mil, shayok@cs.fsu.edu, ftorre@cs.cmu.edu

Abstract

Detecting vehicles in aerial imagery is a critical task with applications in traffic monitoring, urban planning, and defense intelligence. Deep learning methods have provided state-of-the-art (SOTA) results for this application. However, a significant challenge arises when models trained on data from one geographic region fail to generalize effectively to other areas. Variability in factors such as environmental conditions, urban layouts, road networks, vehicle types, and image acquisition parameters (e.g., resolution, lighting, and angle) leads to domain shifts that degrade model performance. This paper proposes a novel method that uses generative AI to synthesize high-quality aerial images and their labels, improving detector training through data augmentation. Our key contribution is the development of a multi-stage, multi-modal knowledge transfer framework utilizing fine-tuned latent diffusion models (LDMs) to mitigate the distribution gap between the source and target environments. Extensive experiments across diverse aerial imagery domains show consistent performance improvements in AP_{50} *over supervised learning on source domain data, weakly supervised adaptation methods, unsupervised domain adaptation methods, and open-set object detectors by 4-23%, 6-10%, 7-40%, and more than 50%, respectively. Furthermore, we introduce two newly annotated aerial datasets from New Zealand and Utah to support further research in this field. Project page is available at: https://humansensinglab.github.io/AGenDA*

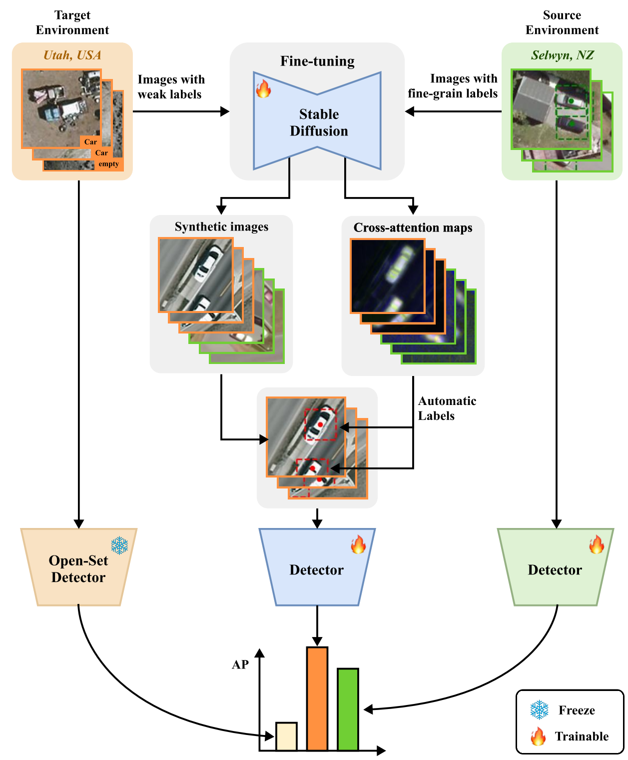


Figure 1. **Overview.** We propose a pipeline that generates high-quality aerial images along with their labels. Our method outperforms baseline detectors trained on source images and open-set detectors directly inferred on target images.

1. Introduction

Recent advancements in Generative AI have led to the development of diffusion-based models [48] that can synthesize images with unprecedented realism, making them nearly indistinguishable from real-world data [11, 34, 50]. Additionally, integrating natural human language has become crucial to modern vision systems. This integration can take the form of conditioning the generation process (for example, in text-to-image generative models [39, 44] utilizing cross-attention mechanism) or serving as an additional modality that shares a common latent space with vision (as seen in models like *CLIP* [40]). One key reason for the success of these new models is that they have been trained on extensive datasets, such as *LAION-5B* [47] or *WIT* [40], consisting of hundreds of millions or billions of text-image pairs. As a result, the scientific community often refers to them as *Foundational Models*.

Upon their inception, researchers have explored how to use foundational models to generate training data for downstream tasks like classification, object detection, and semantic segmentation. Methods such as *ALIA* [7], *DatasetDM* [58], *Dataset Diffusion* [33], *DiffusionEngine* [68] are particularly useful in scenarios with limited or no access to natural training data.

Despite the promising results these methods have achieved augmenting general-purpose datasets like *VOC* [8] and *COCO* [27], little to no effort has been made to generate diverse, high-resolution annotated synthetic datasets on aerial views. These datasets are important for training detectors on small objects like vehicles. The primary reason for this gap is that off-the-shelf generative models, struggle to generate these views due to insufficient aerial imagery representation in their training datasets. Hence, fine-tuning is unavoidable in this case. Recent efforts have primarily produced generative models constrained by geospatial resolution, such as *DiffusionSat* [16] and *SatDiffMoE* [49], and therefore cannot be utilized for vehicle detection.

The lack of sufficient large-scale aerial view training data impacts the performance of not only generative foundational models but also other types of *Vision Large Language Models [65]* (VLLMs), such as *BLIP2* [21], *InternVL3* [73], and *Gemini* [53]. Our analysis reveals that these models perform poorly in zero-shot settings for tasks such as car presence classification and center localization in aerial imagery. (see Appendix Sec. C). This limitation also extends to open-set object detectors [22, 32, 70], which, as shown in Table 1, exhibit unsatisfactory results on aerial imagery.

On the other hand, detection annotations should also be produced for each generated image object. However, text-to-image models do not naturally provide these. There are two potential options to address this challenge: 1) the annotations (*e.g.*, semantic segmentation maps, bounding box layouts) to be created before the image generation and used as a spatial conditioning signal [25, 67], or 2) products from the generative process, such as cross- and self-attention maps extracted from the denoising U-Net [11] to be employed to image produce annotations [33, 52, 58, 68]. In the first case, studies reported that the generative models often do not fully obey the imposed conditioning, especially for small objects such as vehicles in aerial view images [60], which may result in generated annotations that do not match the corresponding synthetic images. Therefore, we studied the prospects of the second approach.

To this end, we introduce a new method for annotated aerial view vehicle detection via synthetic image dataset generation, which employs stacked (multi-channel) cross-attention maps, learnable text prompt tokens, and multi-stage cross-environment (source to target) knowledge transfer. We consider the scenario where a fully annotated dataset with bounding box annotations from the source environment and a target environment dataset with weak binary annotations (whether or not a vehicle is present in a given image) are available, which are much easier to obtain than full bounding box annotations. Our extensive experiments show considerable improvements over baseline (model trained only on the source environment data) performance demonstrated by four popular object detectors.

Our contributions can be summarized as follows:

- We introduce a novel approach for generating aerial view annotated synthetic image datasets based on multi-channel cross-attention maps and multi-stage cross-environment knowledge transfer.
- We conduct extensive experiments involving four popular state-of-the-art object detectors to assess the effectiveness of our approach. We compared the performance of our method with other published methods for open-set detection, unsupervised object detector adaptation, and weakly supervised model adaptation. The results corroborate the promise and potential of our framework.
- We introduce two new real-world aerial view datasets captured in Selwyn (New Zealand), 2,078,077 images, and Utah (USA), 2,684,658 images, including car (small vehicle) detection task annotations.

2. Related Work

2.1. Diffusion Models for Perception Tasks

Diffusion models [5, 11] have undergone significant developments and have emerged as prominent generative models in modern research. These models work by progressively corrupting data with noise, and then learning to reverse this process to reconstruct the original data. Once trained, new data can be generated by applying the learned denoising process to randomly sampled noise. Latent diffusion models (LDM) [39, 44] perform diffusion process in the latent image space, which reduces the computation cost towards high-resolution image synthesis. Customized image generation can be achieved by adding various types of control, such as texts [45], edges [67], segmentation masks [36], geometric layouts [3], and images [41].

Recent research demonstrates that diffusion models can also be utilized in various perception tasks. Diffusion Classifier [20] reveals that pre-trained diffusion models can be directly employed for zero-shot classification tasks. DoGE [57] conditions Stable Diffusion [44] on CLIP image embedding difference between two domains, improving classification and semantic segmentation accuracy. DGInStyle [13] combines semantic masks with style prompts to generate training data for semantic segmentation. DatasetDM [58] uses a few labeled real images to train a mask decoder, leading to a robust synthetic data generator. Thanks to the cross-attention mechanism in Stable Diffusion, several works [33, 56, 59] obtain high-quality segmentation labels through cross-attention maps between images and target concepts. AttnDreambooth [35] further incorporates learnable tokens before target concepts to enhance the accuracy of cross-attention maps. Compared to previous works, our method differs in several aspects. First, we explore Stable Diffusion for cross-domain object detection where diffusion models are rarely trained, while previous works mostly focus on classification and segmentation. Second, we synthesize accurate labels by introducing learnable tokens to encode foreground and background concepts for more accurate cross-attention maps and labeling based on style-less cross-attention maps to remove domain gaps.

2.2. Cross-domain Object Detection

Cross-domain object detection addresses the challenge of detecting objects when a domain shift exists between the source and target environments. It can be categorized into subfields based on the availability of data and labels during the training and testing phases.

Open-set detection is helpful in detecting novel categories in the target domain since it identifies both known and unknown objects during inference. One approach is aligning textual and visual features by jointly training object detection [22]. Another approach includes integrating transformer-based architectures or YOLO framework with open-set object detection, improving robustness in detecting objects without category constraints [4, 29]. Vision transformer-based models [31, 32] further enhance open-vocabulary object detection by utilizing large-scale text-image pre-training, enabling models to detect beyond predefined categories. More recent research [6, 70] refines open-set detection through improved representation learning and domain generalization techniques.

Unsupervised domain adaptation for object detection aims to improve detection performance in an unlabeled target domain by leveraging various techniques. A common approach is style transfer [37, 57, 63], which aligns the visual characteristics of source and target domains to mitigate domain shifts. Recent research also uses knowledge distillation [15, 72], where knowledge from a well-trained teacher model is transferred to a student model, enabling adaptation to the target domain. Other methods include adversarial feature learning [26, 69] which minimizes domain discrepancies through adversarial training, or graph reasoning [23] that captures structural relationships between objects, enhancing robustness in cross-domain object detection.

Weakly supervised domain adaptation enhances object detection in target domains where only limited supervision is available and can be categorized into network-based methods and generative models. Network-based methods improve feature alignment and domain generalization through approaches such as hierarchical feature learning [64], transformer-based adaptations [18], and pseudo-labeling [12]. These methods refine object detection models by utilizing structured architectural modifications that help bridge domain gaps. Generative models [13, 17] enhance adaptation by learning domain-invariant representations through style transfer and task-adaptive pre-training. By synthesizing target-like visual features, these approaches help models generalize more effectively to new environments with limited supervision, mitigating domain discrepancies while preserving task-relevant information.

3. Method

Given the fully annotated (where all vehicles in each image are annotated with bounding boxes) source domain data represented as $\mathbf{D}^{\mathcal{S}} = \left\{\left(x_i^{\mathcal{S}}, y_i^{\mathcal{S}}\right)\right\}_{i=1}^{N_{\mathcal{S}}}$, and image-level annotated (whether or not a vehicle is present in a given image) target domain data $\mathbf{D}^{\mathcal{T}} = \left\{\left(x_i^{\mathcal{T}}, y_i^{\mathcal{T}}\right)\right\}_{i=1}^{N_{\mathcal{T}}}$, the goal is to synthesize a fully annotated dataset $\mathbf{D}^{\mathcal{G}} = \left\{\left(x_i^{\mathcal{G}}, y_i^{\mathcal{G}}\right)\right\}_{i=1}^{N_{\mathcal{G}}}$ in the target domain, where x and y denote the images and their corresponding labels, respectively. $N_{\mathcal{S}}$, $N_{\mathcal{T}}$ and $N_{\mathcal{G}}$ denote the total number of images in each dataset. In Sec. 3.1, we present our approach for generating synthetic images with Stable Diffusion [44] by fine-tuning the model to adapt to both source and target domains. In Sec. 3.2, we leverage multi-channel cross-attention maps for object localization in synthetic images. In Sec. 3.3, we propose an automatic labeling strategy to generate pseudo labels for synthetic target domain data. A summary of the full pipeline is provided in the Appendix Sec. A.

3.1. Text-to-image generation

Stable Diffusion consists of an image encoder $\mathcal{E}$, a conditional U-Net ϵ_θ, a text encoder τ_θ, and a latent decoder $\mathcal{D}$. We fine-tune the U-Net ϵ_θ on both $\mathbf{D}^{\mathcal{S}}$ and $\mathbf{D}^{\mathcal{T}}$. In the forward process, an image x is encoded into a latent representation $z_0 = \mathcal{E}(x)$. A noisy latent z_t at any time t is then sampled using the following sampling function [11]:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon,\ \varepsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{1}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, and t is uniformly sampled from $\{1, ..., T\}$. To learn the reverse process, the latent z_t is passed to the U-Net ϵ_θ, along with the timestep t and the prompt embedding $\tau_\theta(c)$. To encode domain-specific information, we design distinct prompts for source and target domain images, denoted as $c_{\mathcal{S}}$ and $c_{\mathcal{T}}$. $c_{\mathcal{S}}$ and $c_{\mathcal{T}}$ follow the format of “an aerial image with [category] in [S]” and “an aerial image with [category] in [T]”, where [category]

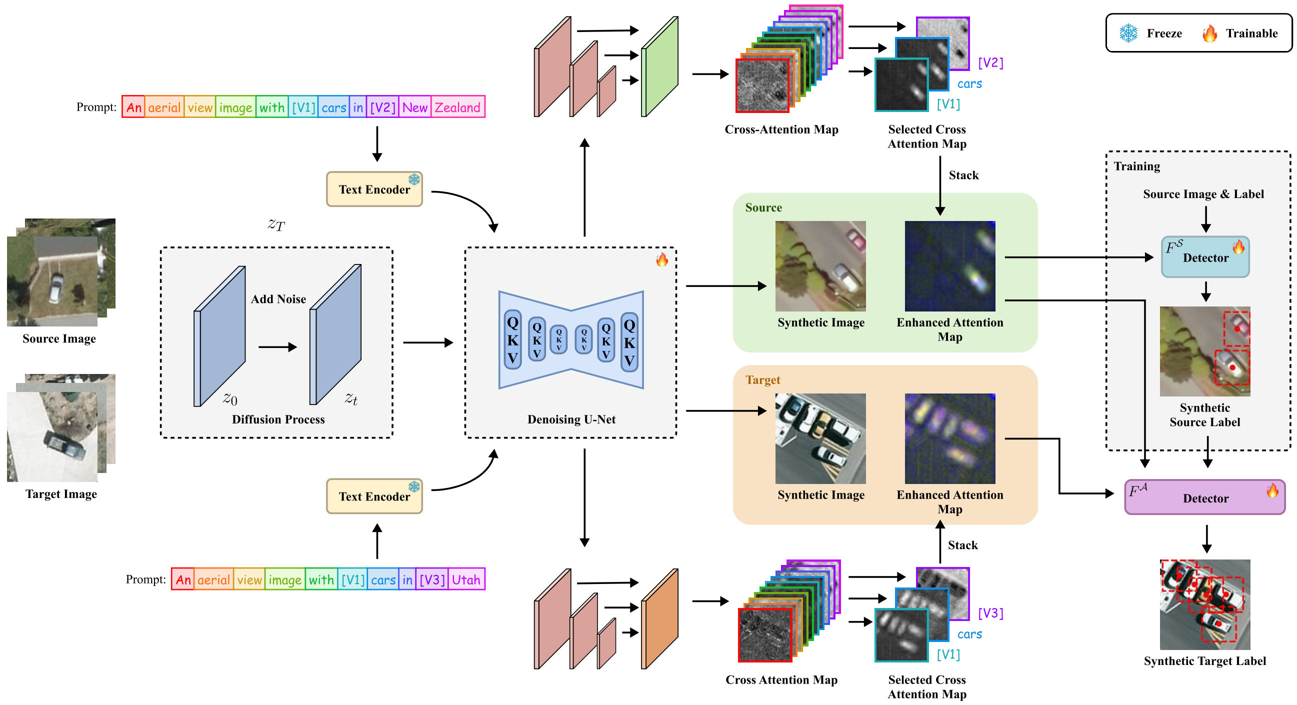


Figure 2. **Overview of our pipeline.** It consists of two stages. First, we finetune Stable Diffusion and synthesize both source and target domain images. Second, we automatically label synthetic target domain images via cross-attention maps.

represents the object type, and [S] and [T] are unique identifiers to distinguish between the source and target domain prompts, as inspired by [45]. The U-Net ϵ_θ is trained to predict the noise added to the latent z_0:

$$L_{\text{LDM}} = \mathbb{E}_{\mathcal{E}(x),c,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(c))\|_2^2\right] \quad (2)$$

where $c \in \{c_\mathcal{S}, c_\mathcal{T}\}$ depends on the domain of the input image x. During image generation, a pure noise latent z_T is iteratively denoised through the U-Net ϵ_θ for T steps, followed by decoding via the latent decoder $\mathcal{D}$ to generate the final image. By conditioning the U-Net ϵ_θ on distinct prompts $c_\mathcal{S}$ and $c_\mathcal{T}$, we can independently synthesize images corresponding to the source and target domains.

3.2. Concept Localization

It has been observed in [52] that in a well-trained text-to-image diffusion model, cross-attention maps between the feature maps and the conditioning text assign higher weights to regions that align with the text concept. This indicates that cross-attention maps effectively help locate the concept. Therefore, we propose leveraging cross-attention maps corresponding to "[category]" in a fine-tuned Stable Diffusion to facilitate downstream visual tasks. Specifically, cross-attention maps can be obtained in each of the four layers in the U-Net ϵ_θ, which corresponds to four different resolutions. We compute the attention map $\mathcal{A}$ by averaging all the cross-attention maps across these resolutions. Finally, we normalize $\mathcal{A}$ to $(0, 1)$ to emphasize regions with higher attention weights.

While cross-attention maps for the object category effectively highlight relevant regions, their robustness can be further improved by cross-verifying with additional maps. Inspired by [9, 35], we introduce two learnable tokens in the prompt to generate complementary cross-attention maps. These tokens are designed to capture both the object and the background concept, which includes all regions outside the objects. The insight is that combining a learnable context token with the object category enhances the localization of the target objects, while a background token helps identify non-target regions. By effectively localizing the background, we can further refine object delineation. To implement this approach, we design a two-stage pipeline. Throughout the two stages, we set the prompt as "an aerial view image with $[V_1]$ [category] in $[V_2]$ [S]" for source domain data and "an aerial view image with $[V_1]$ [category] in $[V_3]$ [T]" for target domain data. $[V_1]$ represents the learnable token for the object concept, while $[V_2]$ and $[V_3]$ correspond to the learnable token for the source domain and target domain background concept, respectively.

In the first stage, we fine-tune both the U-Net ϵ_θ and learnable tokens $[V_1]$, $[V_2]$, and $[V_3]$ to capture the new concepts. To facilitate this process, we introduce a novel cross-attention map regularization loss that encourages similarity between the attention maps of $[V_1]$ and [category], while penalizing similarity between the attention maps of $[V_2]$, $[V_3]$ and [category]. We denote the normalized cross-attention map of $[V_1]$, $[V_2]$, $[V_3]$, and [category] as A_{V_1}, A_{V_2}, A_{V_3} and A_{c}, respectively. To enforce similarity between A_{V_1} and A_{c}, we further normalize them into discrete probability distributions where pixel values sum to one while preserving their relative differences. We denote them as $\hat{A}_{\mathrm{V}_1}$ and $\hat{A}_{\mathrm{c}}$. We then employ the total variation distance metric to minimize the difference between $\hat{A}_{\mathrm{V}_1}$ and $\hat{A}_{\mathrm{c}}$, which is equivalent to half of the L_1 distance between them [19]:

$$L_{\text{obj}} = \frac{1}{2}\sum_{x,y} |\hat{A}_{\mathrm{V}_1}(x, y) - \hat{A}_{\mathrm{c}}(x, y)| \tag{3}$$

Similarly, to penalize similarity between A_{bg} and A_{c} where $A_{\text{bg}} \in \{A_{\mathrm{V}_2}, A_{\mathrm{V}_3}\}$, which is equivalent to enforce similarity between A_{bg} and A^*_{c} where $A^*_{\mathrm{c}} = 1 - A_{\mathrm{c}}$, we first normalize both A_{bg} and A^*_{c} into discrete probability distributions and then compute the distribution difference using the total variation difference metric as follows:

$$L_{\text{bg}} = \frac{1}{2}\sum_{x,y} |\hat{A}_{\text{bg}}(x, y) - \hat{A}^*_{\mathrm{c}}(x, y)| \tag{4}$$

The total loss function can be formulated as $L = L_{\text{LDM}} + L_{\text{obj}} + L_{\text{bg}}$.

In the second stage, we fix the learned tokens $[V_1]$, $[V_2]$, and $[V_3]$, and further fine-tune the U-Net ϵ_θ. This is because Stable Diffusion learns to fit the data distribution with varying prompts in the previous stage, which might not align well with the data distribution conditioned on the final learned prompt. We continue to apply the loss function L to guide the learning of the attention maps and avoid any embedding misalignment.

3.3. Automatic Labeling via Cross-Attention Maps

After fine-tuning Stable Diffusion as described in Sec. 3.1 and Sec. 3.2, we generate synthetic source domain data $\mathbf{D}^{\mathcal{GS}} = \left\{\left(x_i^{\mathcal{GS}}, \widetilde{A}_i^{\mathcal{GS}}\right)\right\}_{i=1}^{N_{\mathcal{GS}}}$ and target domain data $\mathbf{D}^{\mathcal{GT}} = \left\{\left(x_i^{\mathcal{GT}}, \widetilde{A}_i^{\mathcal{GT}}\right)\right\}_{i=1}^{N_{\mathcal{GT}}}$. For each synthetic data sample $D_i \in \mathbf{D}^{\mathcal{GS}} \cup \mathbf{D}^{\mathcal{GT}}$, we extract the cross-attention maps $A_{i,\mathrm{c}}$, A_{i,V_1} and $A_{i,\text{bg}}$ during denoising steps. The enhanced cross-attention map $\widetilde{A}_i$ is then obtained by stacking these components. Since cross-attention maps are grayscale images that highlight object regions, they contain less style information than RGB images. Therefore, we propose using these maps to generate bounding box annotations for target domain data. First, we train a detector $F^{\mathcal{S}}(\cdot;\theta)$ on the fully annotated real source domain data $\mathbf{D}^{\mathcal{S}}$. This detector is then used to predict reliable pseudo labels $\{y_i^{\mathcal{GS}}\}_{i=1}^{N_{\mathcal{GS}}}$ on the synthetic source domain images $\{x_i^{\mathcal{GS}}\}_{i=1}^{N_{\mathcal{GS}}}$. Next, we train another detector $F^{\mathcal{A}}(\cdot;\theta)$ on the combined dataset $\mathbf{D}^{\mathcal{GS}}_* = \left\{\left(\widetilde{A}_i^{\mathcal{GS}}, y_i^{\mathcal{GS}}\right)\right\}_{i=1}^{N_{\mathcal{GS}}}$, which comprises synthetic source domain enhanced cross-attention maps and their corresponding pseudo labels. Finally, we use the well-trained detector $F^{\mathcal{A}}(\cdot;\theta)$ to test the target domain cross-attention maps $\left\{\widetilde{A}_i^{\mathcal{GT}}\right\}_{i=1}^{N_{\mathcal{GT}}}$ and predict a set of pseudo labels $\{y_i^{\mathcal{GT}}\}_{i=1}^{N_{\mathcal{GT}}}$ for synthetic target domain data $\mathbf{D}^{\mathcal{GT}}$. This results in a fully annotated synthetic target domain dataset $\left\{\left(x_i^{\mathcal{GT}}, y_i^{\mathcal{GT}}\right)\right\}_{i=1}^{N_{\mathcal{GT}}}$, which can be directly used for downstream object detection tasks.

Determining the confidence score threshold for bounding box labels is challenging because it can vary across datasets. To address this, we propose a classifier refinement method that selects more reliable labels for synthetic target domain data, inspired by [51]. Predicted bounding boxes with high confidence scores represent foreground objects, while those with low scores represent background regions. We define bounding boxes above a high threshold λ_{high} as positive samples and those below a low threshold λ_{low} as negative samples. We train a classifier using these samples on cropped image patches, refining predictions for samples with intermediate confidence scores, resulting in more reliable labels for synthetic target domain data.

4. Datasets

We use three real-world aerial view vehicle detection datasets for our experiments—the publicly available *DOTA* [62], and two additional datasets we created, called *LINZ* and *UGRC* (Figure 3). All three have *ground sampling distance* (GSD) of 12.5 cm per px and have been sampled to 112 px × 112 px image size. Utilizing this image size is essential for our method because diffusion models are known for struggling with generating small objects caused by the cross-attention mechanism's limited resolution. Thus, we increase the relative object size within the images. We use only one object class (*small vehicle*) and location labels (object centers).

DOTA: We used the *DOTA-v2* dataset's training and validation sets, which contain 1,830 and 593 images, respectively, with 169,268 and 56,062 unique small car instances. The image sizes range from 421 px × 346 px to 29,200 px × 27,620 px. We only used images with GSD of 15 cm per px or less and scale them to match the target 12.5 cm per px. The images with larger GSD or whose metadata do not provide such information are rejected. Finally, we produce our experiments' training and validation sets by placing randomly rotated square sampling windows with size 112 px × 112 px uniformly distributed across each original image area. This results in 455,099 training images (28,453 contain objects) and 142,262 validation images (9,460 contain objects), with 62,178 and 20,194 total object instances,

Figure 3. **Image samples from our datasets.** (*left*) LINZ sample, (*right*) UGRC, (*green markers*) small vehicle location annotations. For more examples, check the Supplementary Material.

respectively. The original bounding box labels were converted to object locations to match the annotations of the other two datasets described below.

LINZ: We created this dataset by manually annotating vehicle locations in aerial images captured in *Selwyn, New Zealand* in December 2012. They were obtained from the *Land Information New Zealand* (LINZ) online platform [28]. The original tiles' size is 240 m $\times$ 360 m (1920 px $\times$ 2880 px) with GSD of 12.5 cm per px. A uniformly distributed sampling window produces the final samples with size 112 px $\times$ 112 px. We finally have 1,451,144 training images (19,564 of which include objects), 188,744 validation images (2,108 of which include objects), and 438,189 test images (2,629 of which include objects). The number of object instances in the training set is 29,495, in the validation set, 2,574, and in the test set, 3,640. The Appendix Sec. B provides more details.

UGRC: We used the same approach as with *LINZ* to construct this dataset. We obtained 40 high-resolution (12.5 cm per px) aerial image tiles captured in *Utah, USA* and annotated the locations of small vehicle instances. The images were downloaded from *Utah Geospatial Resource Center* (UGRC) [55] online platform. Their native size is 2,000 m $\times$ 2,000 m (16,000 px $\times$ 16,000 px). After performing dataset sampling using a 112 px square window, we constructed splits with the following quantities: 2,142,849 training images (15,631 of which include objects), 271,252 validation images (3,912 of which include objects), and 270,557 test images (1,510 of which include objects). The number of object instances in the training set is 26,001, in the validation set 9,878, and in the test set 1,900. More information can be found in the Appendix Sec. B.

5. Experiments

5.1. Implementation Details

Working with location-based annotations: As discussed in Sec. 4, LINZ and UGRC datasets contain small vehicle location labels. DOTA's bounding box annotations were converted to location labels for consistency. To leverage the abundance of bounding box-based open-source object detection frameworks, we reformulate the location detection problem as a bounding box detection task. Specifically, we define a decision circle with a 12 px radius centered on the vehicle's centroid. A predicted center is considered a true positive if it falls within this circle. Then we introduce a 42.36 px square pseudo-bounding box centered at the vehicle's centroid. This specific size is chosen to facilitate the evaluation using the AP_{50} object detection metric. Mathematically, this approach is functionally equivalent to using the aforementioned decision circle, with minimal error.

Automatic Labeling: To establish detection baselines, we employ the MMDetection framework [2], and evaluate a diverse set of models, including Faster-RCNN [43], YOLOv5 [14], YOLOv8 [42] and ViTDet [24], which represent one-stage, two-stage and transformer-based object detectors. The automatic labeling process follows the methodology outlined in section 3.3, ensuring consistency by utilizing the same detector throughout the entire process. For label refinement, we fine-tune a pre-trained ResNet [10] on the ImageNet [46] dataset for 80 epochs with a batch size of 256. The predefined confidence thresholds λ_{high} and λ_{low} are determined based on the confidence score that yields the optimal F1-score for synthetic source domain cross-attention maps during detector training. Specifically, we set $\lambda_{\text{high}} = 0.7$ and $\lambda_{\text{low}} = 0.3$ for both YOLOv5 and YOLOv8, while for ViTDet and Faster-RCNN, we assign values of $\lambda_{\text{high}} = 0.95$ and $\lambda_{\text{low}} = 0.5$.

Image Generation: We employ Stable diffusion V1.4 [44], pre-trained on the images of size 512×512, a batch size of 64, and a learning rate of 10^{-6} on two RTX A6000 GPUs for approximately 15 epochs. Following section 3.2, we adopt a two-stage fine-tuning strategy to capture both object and background concepts. For the source domain, the guidance prompt is "An aerial view image with [V_1] cars in [V_2][id]" for images with small vehicles since most small vehicles fall under the broader classification of cars (see Appendix Sec. B for more details), and "An aerial view image in [V_2][id]" for images without small vehicles, where [id] is replaced with "New Zealand" or "DOTA". For the target domain, [V_2] is substituted with token [V_3] to learn unique background concept in target domain, and [id] is replaced with "Utah". In the first stage, we fine-tune the learnable tokens in the prompt along with the U-Net. In the second stage, we freeze the learned tokens and fine-tune U-Net to better align the model with the data distribution. Each stage is trained for approximately two epochs using a batch size of 8 and a learning rate of 5×10^{-7} on two RTX A6000 GPUs. We synthesize 10*k* images containing cars for both the source and target domain, and 10*k* images without cars for the target domain. Additional implementation details are provided in the Appendix Sec. D.

Method	Backbone	LINZ→UGRC $AP_{50}(\%)$	DOTA→UGRC $AP_{50}(\%)$
Supervised Object detection (Source-only)			
Faster-RCNN [43]	Faster R-CNN [43]	53.1	52.5
YOLOv5 [14]	YOLOv5 [14]	63.2	57.6
YOLOv8 [14]	YOLOv8 [42]	62.9	71.4
ViTDet [24]	ViTDet [24]	55.7	50.4
Open-set object detection			
GLIP-T [22]	Swin-T [30]	8.7	8.7
OmDet-Turbo [70]	Swin-T [30]	14.4	14.4
OWLV2 [32]	CLIP ViT [40]	17.9	17.9
Unsupervised domain adaptation object detection			
SIGMA [23]	Faster R-CNN [43]	36.2	34.8
TIA [69]	Faster R-CNN [43]	49.2	42.4
Adapt. Teacher [26]	Faster R-CNN [43]	29.0	37.0
CycleGAN-Turbo [37]	YOLOv5 [14]	56.0	60.8
SSDA-YOLO [72]	YOLOv5 [14]	52.3	49.6
Cross domain weakly supervised object detection			
OCUD [71]	Faster R-CNN [43]	63.1	65.3
H2FA R-CNN [64]	Faster R-CNN [43]	61.8	68.3
Ours	Faster R-CNN [43]	**69.3**	**75.5**
Ours	YOLOv5 [14]	**68.8**	**68.5**
Ours	YOLOv8 [42]	**75.4**	**75.7**
Ours	ViTDet [24]	**72.0**	**67.1**

Table 1. **Cross-domain object detection results**. LINZ to UGRC and DOTA to UGRC. We report the AP_{50} result.

5.2. Comparison with State-of-the-art Methods

Table 1 summarizes the cross-domain detection results for adaptation from the LINZ and DOTA datasets to the UGRC dataset. First, our method consistently outperforms detectors trained solely on source domain data, demonstrating its robustness and generalizability. Second, open-set detectors perform poorly on aerial images, highlighting the lack of training of foundational models in this domain. Third, compared to domain adaptation methods [23, 26, 64, 69, 71] based on Faster-RCNN [43], our pipeline achieves the highest AP_{50} of 75.4% for LINZ → UGRC, and 75.7% for DOTA → UGRC, surpassing the best unsupervised method by 20.1% and 33.1%, and the best weakly supervised method by 6.2% and 7.2%, respectively. Similarly, using YOLOv5 [14] as the backbone, our pipeline outperforms existing methods [37, 72] by 12.8% and 7.7%. These results strongly corroborate the promise and potential of our method against competing baselines, for the challenging task of vehicle detection from aerial imagery.

5.3. Ablation Studies

5.3.1. Effectiveness of learnable tokens

We evaluate the effectiveness of stacking three cross-attention maps of the word “car”, learnable tokens of car concept and background for small vehicle detection using YOLOv5 and YOLOv8, as summarized in Table 2.

Method	Backbone	LINZ→UGRC $AP_{50}(\%)$	DOTA→UGRC $AP_{50}(\%)$
Base	YOLOv5	62.1	68.0
Base	YOLOv8	69.8	74.5
Stack	YOLOv5	68.8	68.5
Stack	YOLOv8	75.4	75.7

Table 2. **Comparison between labeling using single cross-attention map and multi-channel cross-attention maps**. “Base” denotes using cross-attention maps of word “car” only. “Stack” denotes stacked cross-attention maps extracted from the word “car” and two learnable tokens for the car concept and background.

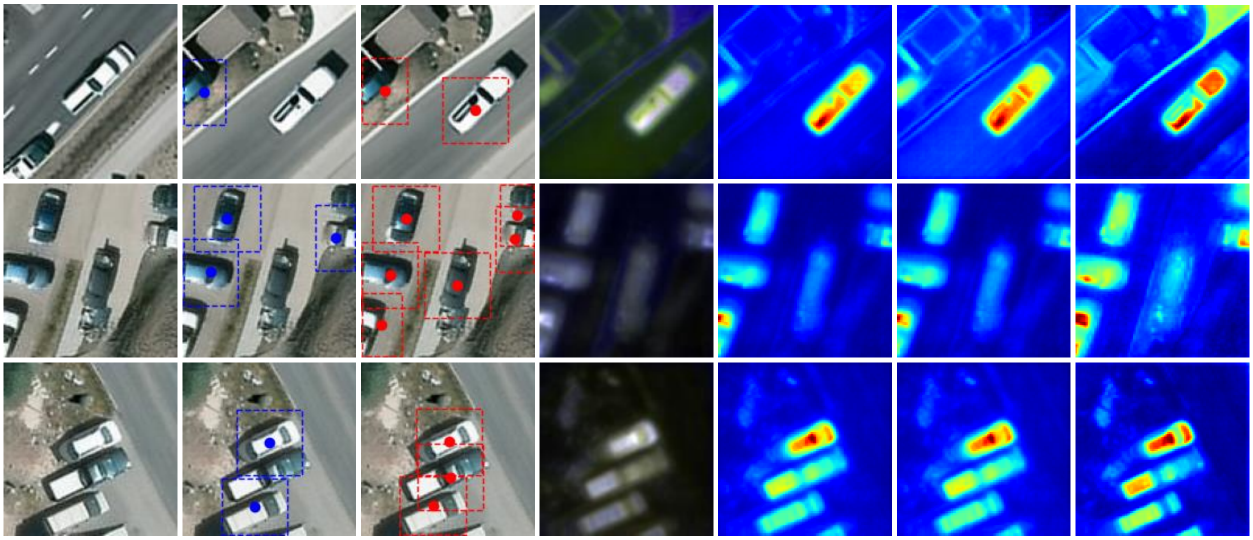
Compared to using only the cross-attention map of the word “car”, incorporating multiple cross-attention maps improves detection accuracy by 5.6% when adapting from LINZ to UGRC, and by 1.2% from DOTA to UGRC. Additionally, we provide visualizations of cross-attention maps corresponding to different tokens and analyze their impact on labeling synthetic images, as illustrated in Figure 4. The visualizations reveal that different cross-attention maps tend to focus on distinct regions of the cars. By cross-verifying using multiple maps, the labeling process becomes more robust, leading to improved label quality.

5.3.2. Effectiveness of label refinement

To evaluate the effectiveness of fine-tuning a classifier to refine labels, we compare our method against a fixed filtering confidence threshold set at 0.3, 0.4, 0.5, 0.6, and 0.7 using YOLOv8. As shown in Figure 5, the detection performance varies across different datasets without a consistent pattern about the confidence threshold. When compared to the results obtained using different thresholds, our method demonstrates competitive performance, suggesting that it produces a reliable set of labels. Additionally, compared to our method, when setting a lower threshold the detector labels more false positive samples, while setting a higher threshold the model labels fewer cars, as shown in Figure 6. This indicates our method can be helpful to select a more reliable set of labels.

6. Conclusions

This paper introduces a novel approach that leverages diffusion models to automatically generate synthetic aerial view images alongside object location annotations based on multi-channel cross-attention maps. Extensive experiments conducted with four object detectors, as well as comparisons with open-set detectors, unsupervised domain adaptation methods, and weakly supervised models, demonstrate the effectiveness of our method in object detection for aerial imagery. Additionally, we introduce two new aerial view datasets from New Zealand and Utah, including annotations for small vehicle detection tasks. For future work, we plan



(a) (b) (c) (d) (e) (f) (g)

Figure 4. **Cross-attention maps for different tokens.** We analyze the effectiveness of utilizing multi-channel cross-attention maps by assessing the label quality synthetic UGRC images. (a) Synthetic UGRC images. (b) Labels generated using only the cross-attention map of the word "car". (c) Labels generated by using multi-channel heatmaps. (d) Multi-channel cross-attention maps. (e) Cross-attention maps of the word "car". (f) Cross-attention maps of token [V_1], which is designed to capture the concept of cars. (g) Cross-attention maps of token [V_3] for background concept, which are further inverted for better comparison. In (b) and (c), bounding boxes with dotted lines denote the predicted pseudo-bounding box labels while dots denote predicted vehicle centers. In (e), (f), and (g), grayscale cross-attention maps are displayed as color heatmaps to highlight the intensity difference.

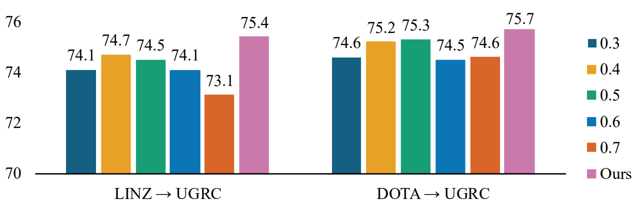


Figure 5. **Quantitative comparison with varying thresholds.** We report the AP_{50} result.

(a) (b) (c) (d) (e)

Figure 6. **Comparison of our method with different thresholds for label quality.** Blue bounding boxes represent pseudo labels generated using fixed thresholds, while red bounding boxes correspond to labels obtained through refinement. The blue and red dots indicate the predicted car centers. (a) to (e) show the results for thresholds of 0.3, 0.4, 0.5, 0.6, and 0.7, respectively.

to integrate VLLMs trained for aerial view images to automatically identify the presence of vehicles in aerial imagery. This would allow us to generate weak labels within the pipeline itself, enabling a fully unsupervised workflow.

Despite its strengths, our method has certain limitations that should be considered. First, we sample large aerial view images into smaller patches of resolution 112 px $\times$ 112 px to mitigate challenges in detecting small objects. This limitation arises because Stable Diffusion progressively downsamples the cross-attention map to a size of 8×8. Consequently, target objects may occupy less than a 1×1 grid, making it difficult for the model to generate precise attention maps. Second, our approach may encounter difficulties in handling overlapping objects. When multiple objects are close together, their cross-attention maps overlap, making it challenging for detectors to separate individual objects when using styleless cross-attention maps for synthetic data annotations.

7. Acknowledgement

We thank Jessica Hodgins, Justin Macey, Melanie Danver, Katie Tender, and Eric Yu for their valuable help in creating the datasets. This work has been funded by the DEVCOM Army Research Laboratory.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2

[2] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 6

[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. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[4] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. YOLO-World: Real-Time Open-Vocabulary Object Detection. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[5] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*, 2021. 2

[6] Penghui Du, Yu Wang, Yifan Sun, Luting Wang, Yue Liao, Gang Zhang, Errui Ding, Yan Wang, Jingdong Wang, and Si Liu. LaMI-DETR: Open-Vocabulary Detection with Language Model Instruction. In *Proceedings of the European conference on computer vision (ECCV)*, 2024. 3

[7] Lisa Dunlap, Alyssa Umino, Han Zhang, Jiezhi Yang, Joseph E Gonzalez, and Trevor Darrell. Diversify your vision datasets with automatic diffusion-based augmentation. In *Advances in Neural Information Processing Systems*, pages 79024–79034. Curran Associates, Inc., 2023. 2

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012. 2

[9] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *The Eleventh International Conference on Learning Representations*, 2023. 4

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, pages 6840–6851. Curran Associates, Inc., 2020. 1, 2, 3

[12] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5001–5009, 2018. 3

[13] Yuru Jia, Lukas Hoyer, Shengyu Huang, Tianfu Wang, Luc Van Gool, Konrad Schindler, and Anton Obukhov. DGInStyle: Domain-Generalizable Semantic Segmentation with Image Diffusion Models and Stylized Semantic Control. In *Computer Vision – ECCV 2024*, pages 91–109, Cham, 2025. Springer Nature Switzerland. 3

[14] Glenn Jocher. Yolov5 by ultralytics, 2020. 6, 7, 2

[15] Justin Kay, Timm Haucke, Suzanne Stathatos, Siqi Deng, Erik Young, Pietro Perona, Sara Beery, and Grant Van Horn. Align and distill: Unifying and improving domain adaptive object detection. *arXiv preprint arXiv:2403.12029*, 2024. 3

[16] Samar Khanna, Patrick Liu, Linqi Zhou, Chenlin Meng, Robin Rombach, Marshall Burke, David B. Lobell, and Stefano Ermon. DiffusionSat: A Generative Foundation Model for Satellite Imagery. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[17] Neehar Kondapaneni, Markus Marks, Manuel Knott, Rogerio Guimaraes, and Pietro Perona. Text-Image Alignment for Diffusion-Based Perception. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13883–13893, 2024. 3

[18] Tyler LaBonte, Yale Song, Xin Wang, Vibhav Vineet, and Neel Joshi. Scaling novel object detection with weakly supervised detection transformers. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 85–96, 2023. 3

[19] David A Levin and Yuval Peres. *Markov chains and mixing times*. American Mathematical Soc., 2017. 5

[20] Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2206–2217, 2023. 2

[21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 2, 3

[22] Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022. 2, 3, 7

[23] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma: Semantic-complete graph matching for domain adaptive object detection. In *CVPR*, 2022. 3, 7

[24] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *Computer Vision – ECCV 2022*, pages 280–296, Cham, 2022. Springer Nature Switzerland. 6, 7, 2

[25] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22511–22521, 2023. 2, 3

[26] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-Domain Adaptive Teacher for Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 7

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence 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*, pages 740–755. Springer, 2014. 2

[28] LINZ. Selwyn 0.125m Urban Aerial Photos (2012-2013). https://data.linz.govt.nz/layer/51926-selwyn-0125m-urban-aerial-photos-2012-2013/, 2014. 6

[29] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying DINO with Grounded Pre-training for Open-Set Object Detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 3, 2

[30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 7

[31] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection. In *Computer Vision – ECCV 2022*, pages 728–755, Cham, 2022. Springer Nature Switzerland. 3, 2

[32] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling Open-Vocabulary Object Detection. In *Advances in Neural Information Processing Systems*, pages 72983–73007. Curran Associates, Inc., 2023. 2, 3, 7

[33] Quang Ho Nguyen, Truong Tuan Vu, Anh Tuan Tran, and Khoi Nguyen. Dataset Diffusion: Diffusion-based Synthetic Data Generation for Pixel-Level Semantic Segmentation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2, 3

[34] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 1

[35] Lianyu Pang, Jian Yin, Baoquan Zhao, Feize Wu, Fu Lee Wang, Qing Li, and Xudong Mao. AttnDreamBooth: Towards Text-Aligned Personalized Text-to-Image Generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3, 4

[36] Dong Huk Park, Grace Luo, Clayton Toste, Samaneh Azadi, Xihui Liu, Maka Karalashvili, Anna Rohrbach, and Trevor Darrell. Shape-Guided Diffusion With Inside-Outside Attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4198–4207, 2024. 2

[37] Gaurav Parmar, Taesung Park, Srinivasa Narasimhan, and Jun-Yan Zhu. One-step image translation with text-to-image models. *arXiv preprint arXiv:2403.12036*, 2024. 3, 7

[38] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[39] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2, 7, 3

[41] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2, 6

[42] Dillon Reis, Jordan Kupec, Jacqueline Hong, and Ahmad Daoudi. Real-time flying object detection with yolov8. *arXiv preprint arXiv:2305.09972*, 2023. 6, 7, 2

[43] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2015. 6, 7, 2

[44] 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 (CVPR)*, pages 10684–10695, 2022. 1, 2, 3, 6, 7

[45] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510, 2023. 2, 4

[46] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 6

[47] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 2

[48] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015. 1

[49] Bowen Song, Zhaoxu Luo, and Liyue Shen. Satdiffmoe: A mixture of estimation method for satellite image super-resolution with latent diffusion models. In *ICML 2024 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2024. 2

[50] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 1

[51] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple Instance Detection Network With Online Instance Classifier Refinement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5

[52] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting Stable Diffusion Using Cross Attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5644–5659, Toronto, Canada, 2023. Association for Computational Linguistics. 2, 4

[53] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2

[54] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 2

[55] UGRC. Utah High Resolution Orthophotography (HRO) 2012 Images. https://gis.utah.gov/products/sgid/aerial-photography/high-resolution-orthophotography/, 2012. 6

[56] Jinglong Wang, Xiawei Li, Jing Zhang, Qingyuan Xu, Qin Zhou, Qian Yu, Lu Sheng, and Dong Xu. Diffusion model is secretly a training-free open vocabulary semantic segmenter. *CoRR*, abs/2309.02773, 2023. 3

[57] Yinong Oliver Wang, Younjoon Chung, Chen Henry Wu, and Fernando De la Torre. Domain Gap Embeddings for Generative Dataset Augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28684–28694, 2024. 2, 3

[58] Weijia Wu, Yuzhong Zhao, Hao Chen, Yuchao Gu, Rui Zhao, Yefei He, Hong Zhou, Mike Zheng Shou, and Chunhua Shen. DatasetDM: Synthesizing Data with Perception Annotations Using Diffusion Models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2, 3

[59] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. DiffuMask: Synthesizing Images with Pixel-level Annotations for Semantic Segmentation Using Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1206–1217, 2023. 3

[60] Yiming Wu, Qihe Pan, Zhen Zhao, Zicheng Wang, Sifan Long, and Ronghua Liang. SOEDiff: Efficient Distillation for Small Object Editing. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2025. 2

[61] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. 2

[62] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5

[63] Sihan Xu, Ziqiao Ma, Yidong Huang, Honglak Lee, and Joyce Chai. CycleNet: Rethinking Cycle Consistency in Text-Guided Diffusion for Image Manipulation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3

[64] Yunqiu Xu, Yifan Sun, Zongxin Yang, Jiaxu Miao, and Yi Yang. H2FA R-CNN: Holistic and Hierarchical Feature Alignment for Cross-Domain Weakly Supervised Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14329–14339, 2022. 3, 7

[65] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12), 2024. 2

[66] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 2

[67] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847, 2023. 2

[68] Manlin Zhang, Jie Wu, Yuxi Ren, Ming Li, Jie Qin, Xuefeng Xiao, Wei Liu, Rui Wang, Min Zheng, and Andy J Ma. Diffusionengine: Diffusion model is scalable data engine for object detection. *arXiv preprint arXiv:2309.03893*, 2023. 2

[69] Liang Zhao and Limin Wang. Task-Specific Inconsistency Alignment for Domain Adaptive Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14217–14226, 2022. 3, 7

[70] Tiancheng Zhao, Peng Liu, Xuan He, Lu Zhang, and Kyusong Lee. Real-time Transformer-based Open-Vocabulary Detection with Efficient Fusion Head. *arXiv preprint arXiv:2403.06892*, 2024. 2, 3, 7

[71] Yuanyi Zhong, Jianfeng Wang, Jian Peng, and Lei Zhang. Boosting weakly supervised object detection with progressive knowledge transfer. In *Computer Vision – ECCV 2020*, pages 615–631, Cham, 2020. Springer International Publishing. 7

[72] Huayi Zhou, Fei Jiang, and Hongtao Lu. SSDA-YOLO: Semi-supervised domain adaptive YOLO for cross-domain object detection. *Computer Vision and Image Understanding*, 229:103649, 2023. 3, 7

[73] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 2