Robust Disaster Assessment from Aerial Imagery Using Text-to-Image Synthetic Data

Tarun Kalluri[1]* **Jihyeon Lee**[2] **Kihyuk Sohn**[2] **Sahil Singla**[2]
Manmohan Chandraker[1] **Joseph Xu**[2] **Jeremiah Liu**[2]
[1]UC San Diego [2]Google Research

Abstract

We present a simple and efficient method to leverage emerging text-to-image generative models in creating large-scale synthetic supervision for the task of damage assessment from aerial images. While significant recent advances have resulted in improved techniques for damage assessment using aerial or satellite imagery, they still suffer from poor robustness to domains where manual labeled data is unavailable, directly impacting post-disaster humanitarian assistance in such under-resourced geographies. Our contribution towards improving domain robustness in this scenario is two-fold. Firstly, we leverage the text-guided mask-based image editing capabilities of generative models and build an efficient and easily scalable pipeline to generate thousands of post-disaster images from low-resource domains. Secondly, we propose a simple two-stage training approach to train robust models while using manual supervision from different source domains along with the generated synthetic target domain data. We validate the strength of our proposed framework under cross-geography domain transfer setting from xBD and SKAI images in both single-source and multi-source settings, achieving significant improvements over a source-only baseline in each case.

1. Introduction

In this work, we address the issue of poor robustness caused by traditional training methods for the task of disaster assessment by generating synthetic data using guided text-to-image generation [8, 32]. To accelerate rescue, recovery and aid routing through scalable and automated disaster assessment from images, recent methods propose efficient training paradigms using paired labeled data from before and after the disaster [5, 16, 28, 50, 52, 61]. While being instrumental in significantly improving the accuracy in damage assessment, these methods greatly rely on manual supervision for efficient performance and perform poorly when deployed in novel domains - such as new disaster types or unseen geographies. While unsupervised adaptation methods exist to overcome the overhead of manual annotation [4, 6, 24], they still require unlabeled images captured from both before and after the disaster for learning domain agnostic features. While readily available satellite imagery provides generalized aerial coverage for most geographic locations for pre-disaster images, the retrieval of post-disaster imagery remains a time-consuming process, hindering rapid damage assessment during critical response windows, with non-trivial domain shifts preventing cross-geographical deployment.

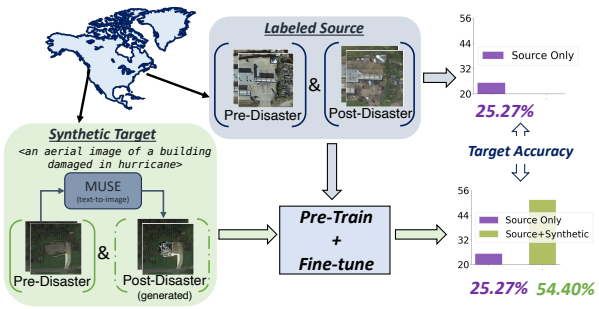


Figure 1. **Summary of our proposed pipeline.** A disaster assessment model trained using labeled data from a different domain suffers from poor accuracy due to significant distribution shifts with the target. We offer a novel way of addressing this limitation, by leveraging the recent advances in mask-based text-to-image models [8] to generate thousands of synthetic labeled data from the target domain where only pre-disaster images are accessible. We incorporate this synthetic data along with source labeled data in a two-stage training framework to achieve significant gains on challening transfer settings from xBD [16] and SKAI [23] datasets.

On the other hand, there has been a notable progress in the field of generating synthetic data using guided text-to-image models [12, 38, 44, 63], which overcome the cumbersome manual annotation process and enable controllable data-generation at scale to train robust and data-efficient models. While a majority of these works focus on tasks like

*Work done during TK's internship at Google.

image recognition [38, 43, 44], object detection [14] and semantic segmentation [29, 57], extending these methods to suit the current setting of aerial disaster assessment is non-trivial, as it requires generating precisely synchronized pre- and post-disaster imagery of the affected area.

We present a novel framework that leverages image-guided text-conditioned generative models to synthesize large datasets of post-disaster imagery conditioned on pre-disaster imagery, which is trained efficiently using a two-stage approach by incorporating unlabeled target domain data alongside source labels. By exploiting the mask-based image-editing abilities of transformer-based text-to-image models, we edit the pre-disaster image using a suitable text-prompt to create a corresponding synthetic post-disaster image, generating a new, synthetically labeled dataset specific to the target domain. To mitigate performance degradation caused by domain shift between generated data and real-world images, we adopt a two-stage training procedure. We first train a siamese vision-transformer [9] using labeled data from the source domains, and subsequently fine-tune the last layer on the synthetic data from the target domain following prior work [21]. We show the effectiveness of our framework in training robust models through experiments on several challenging transfer settings from xBD and SKAI datasets, significantly outperforming a source-only training baseline in each case. As shown in Fig. 1, while training directly using the source domain only achieves only 25% accuracy on the target test-data, our synthetic-data augmented training achieves 54.4%, with a non-trivial improvement of 29% on the challenging xBD dataset. In summary, our contributions are as follows.

1. We offer a cost-effective way to generate training data for disaster assessment in areas lacking real-world aerial imagery, leveraging the image-editing abilities of large-scale text-to-image models.
2. Following prior work in robustness studies [21], we devise a simple and effective two-stage training strategy to use the synthetically generated data in training along with labeled data from different source domains to achieve complementary benefits.
3. We validate the effectiveness of our proposed framework on two benchmark datasets xBD [16] and SKAI [23] images, obtaining significant improvements over a standard source-only baseline in both single-source (+9.8%, +25.2%) and multi-source (+5.33%, +29.13%) domain transfer settings.

2. Related Work

Disaster Assessment using Satellite Images The task of image-based disaster assessment involves predicting the presence or extent of damage in a particular location by comparing pre and post disaster aerial or satellite imagery. Fueled by the availability of paired pre- and post-disaster images captured from remote-sensing satellites [16, 23, 28, 31, 41, 53], several methods have been proposed to identify the damage [2, 9, 26, 49], as well as to precisely localize the damage within the image [15, 30, 51]. However, these approaches rely on labeled data for efficient performance, preventing their use in novel domains without incurring additional collection and annotation overheads [6, 55]. While domain adaptation methods exist to bridge this gap [4, 24, 46], they still need to access the post-disaster imagery which is difficult to acquire in a short window following a disaster. While prior works attempt generation of images using GANs [48], they lack the ability to generate controllable synthetic data at scale. Our work addresses these limitations by leveraging the advances in conditional text-to-image capabilities to generate large-scale synthetic supervision from low-resource target domains. We also note that while there has been significant advances in unsupervised domain adaptation for image classification [19, 37, 56, 64] and segmentation [22, 45], they are typically not applicable to expert tasks like disaster assessment through aerial imagery, preventing their direct use or comparison for our problem.

Creating Synthetic Data from Generative Models Recent progress in the field of generative modeling has enabled the creation of diverse and realistic images conditioned on a variety of inputs such as text [8, 32, 34, 36], images [27, 35], layouts [62] or semantic maps [47, 60]. In particular, text-to-image synthesis enables creation of diverse visual content based on natural language prompts [8, 20, 32, 36, 58]. Recent works explored the use of leveraging the power of these models in generating synthetic supervision for various tasks including object recognition [1, 38, 43, 44, 63], object detection [14, 25, 54], semantic segmentation [29, 57], outlier detection [12] and long-tailed robustness [39, 40, 59]. Building upon this line of work, our work tackles image generation of post-disaster imagery in low-resource domains through localized editing of the corresponding pre-disaster images guided by suitable text prompts, showing an efficient way to improve domain robustness.

3. Method

3.1. Problem Setting

We now describe our problem setting of investigating domain robustness in image-based disaster assessment tasks. We denote our labeled source dataset as $\mathcal{D}_s = \{U_s^i, V_s^i, y^i\}_{i=1}^{N_s}$, where U^i is the *before* image (image captured before the disaster, also called pre-image), V^i is the *after* image (image captured after the disaster, also called the post-image) along with a binary label $y^i \in \{0, 1\}$ indicating whether or not there is damage between the images due to a disaster. N_s denotes the number of source images. In general, the pre and post images from the source images are paired, where

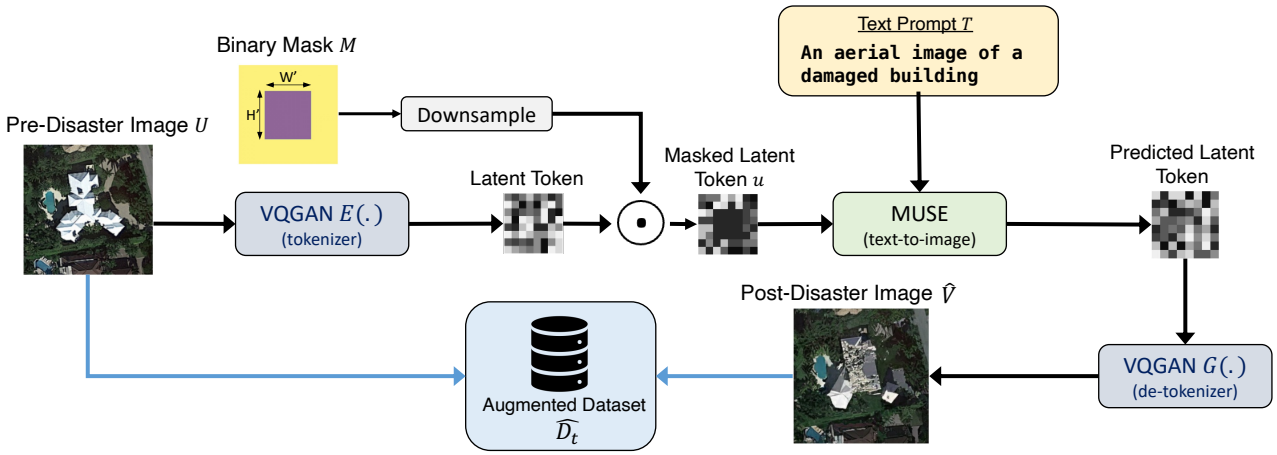


Figure 2. **Overview of the proposed synthetic data generation pipeline.** We first pass the pre-disaster image U from the target domain through a pre-trained VQGAN encoder followed by a tokenizer to compute the latent token, which is then masked using a binary mask. We use the MUSE model along with a suitable text prompt T to predict the output tokens from the masked tokens u, which are then de-tokenized to generate the post-disaster image $\hat{V}_t$. Our augmented dataset $\hat{D}_t$ now contains the input image U_t, generated image $\hat{V}_t$ and the binary label corresponding to the text prompt (indicating damage or no-damage).

the image collection is synchronized to capture the same location before and after the disasters, ensuring pixel-level correspondence between the images. Furthermore, the target domain is denoted by $\mathcal{D}_t = \{U_t^i\}_{i=1}^{N_t}$, where N_t denotes the number of (unlabeled and unpaired) target images. Unique to our work, we assume a *zero-shot* target setting, where we only have access to the pre-images from a new domain (which could be a new geographical location or new disaster type), and neither the post images nor the damage labels are available. This setting is more realistic as capturing paired images immediately after a disaster and labeling them for damage incurs expensive time and manual annotation overhead in most cases, while pre-disaster images are naturally available through satellite footage. Therefore, our main goal is to utilize labeled source images along with unlabeled pre-disaster target images to improve the performance on the target domain at test-time.

While unsupervised domain adaptation (UDA) [4, 24] has been a standard approach to address such domain shifts for disaster assessment, UDA methods still require access to paired pre and post-disaster data during training from both source and target domains. In contrast, we only require images captured before the disaster, and automatically generate synthetic post-disaster images to facilitate robust training.

3.2. Background: Text-to-Image Generation

In our work, we adopt the MUSE [8] model for text-to-image generation, and we provide a brief overview of that framework here. MUSE is a non-autoregressive model for text-conditioned synthesis capable of generating high-resolution images with fast inference speeds. In contrast to diffusion models requiring sequential decoding over several time-steps [10, 32, 34, 36], MUSE adopts a purely transformer-based generation approach outlined in MaskGIT [7] with improved inference speed enabled by parallel decoding. In particular, MUSE uses a pre-trained T5-XXL text encoder [33] to first encode the text-prompt into a 4096-dimensional language embedding, while the input images are passed through a semantic tokenizer such as VQGAN [13] consisting of an encoder and a quantization layer to map the image into a sequence of discrete tokens from a learned codebook, along with a decoder which maps the predicted or generated tokens back into the images. To enable high resolution image synthesis, MUSE adopts the use of two VQGAN models with different downsampling ratio and spatial resolution of the tokens, along with two transformer modules called *base* and *super-resolution* modules that generate low resolution and high resolution latent tokens respectively. During training, these tokens are trained using cross-entropy loss with a reference image, while during inference, the high resolution latent tokens are passed through the VQGAN decoder to generate images conditioned on the input text prompts. We refer the reader to [8] for further details about the training and inference of MUSE generative models.

The use of MUSE in our pipeline as opposed to diffusion-based or auto-regressive based generative models is motivated by two favorable properties. Firstly, MUSE allows efficient mask-based editing capabilities based on its tokenized encoding and decoding-based architecture. Furthermore, MUSE is capable of high-resolution image synthesis with fast inference time and flexible latent token sizes, facilitating the generation of thousands of high quality synthetic

post-disaster images conditioned on pre images at scale.

3.3. Generating Synthetic Data

We provide an overview of our proposed generation pipeline in Fig. 2. We first take the pre-disaster image from the target domain $U_t \in \mathbb{R}^{H\times W\times 3}$, where H, W are the height and width of the image, and generate a corresponding binary mask $\mathcal{M} \in \mathbb{R}^{H\times W}$ with a center patch of size H', W' as ones and rest as zeros, with $H' < H$ and $W' < W$. Since the aerial images in the datasets we considered are typically centered around the subject, we align the center of the binary mask with the center of the image, helping us in directly editing the relevant subject in the image. In practice, we randomly perturb the mask around the center by a factor (δ_x, δ_y) in both dimensions, where the perturbation factor $\delta_x \sim \mathbb{U}[-W/16, W/16]$ and $\delta_y \sim \mathbb{U}[-H/16, H/16]$ is sampled from the uniform distribution. So our patch with all ones in the binary mask is centered around $(H/2 + \delta_y, W/2 + \delta_x)$ with respect to the input image U_t and mask $\mathcal{M}$.

We then pass the target image U_t through the VQGAN encoder to compute the latent tokens for the image, and downsample the binary mask $\mathcal{M}$ to match the spatial resolution of the latent tokens (which is $1/16^{\text{th}}$ and $1/8^{\text{th}}$ of the original image size for the low-resolution and high-resolution pipelines respectively). Subsequently, we multiply the downsampled binary mask $\mathcal{M}'$ with the latent tokens at each pixel location to get a masked token embedding u_t. The masked-latent token embeddings along with the text-embedding of the input prompt are then passed through a series of cross-attention layers pre-trained in MUSE to predict the output tokens from the predicted latent tokens. The outputs are then passed through the decoder layer of the VQGAN resulting in the output image $\hat{V}_t$.

We generate four output images for each pre-image and prompt pair, and pick the best one using the ranking obtained by CLIP similarity score [17] between the input prompt and the generated image. We repeat this process for every image in the target dataset, creating a synthetic dataset $\hat{\mathcal{D}}_t = \{U_t^i, \hat{V}_t^i, \hat{y}^i\}_{i=1}^{N_t}$ of pre and post disaster images from the target domain.

Prompt Pool for Generation A major advantage of generating synthetic images is avoiding the need for manual annotation for unlabeled domains, as the labels can be directly derived from the corresponding prompts. In our setting with binary labels indicating damaged or not damaged buildings or locations, we choose the prompts to reflect these criterion. For example, to create synthetic images for scenes damaged by *hurricane* disaster, we use prompts such as *An aerial view of a house damaged due to a hurricane* or *A satellite image of a building that was destroyed by a hurricane* and assign the label 1. Alternatively, for generating images which have no damage, we create a prompt pool indicating images which are undamaged (for example, *A satellite image of a building*) and assign the generated images with a label of 0. We list the pool of prompts adopted in our work in Fig. 3 and Fig. 4.

SKAI Dataset

Damaged Set
An aerial view of a house damaged due to a hurricane.
A bird's-eye view of a building destroyed by a hurricane.
A top-down view of a house damaged by a hurricane.
A satellite image of a building destroyed by a hurricane.
A bird's-eye view of a building damaged by a hurricane.

Undamaged Set
A satellite image of a house covered by trees.
A bird's-eye view of a house surrounded by trees.
A top-down view of a house under tree shade.
An aerial view of an intact house under tree shade.

Figure 3. Prompt Pool for SKAI

xBD Dataset

Moore Tornado
An aerial view of a house damaged due to a tornado.
A bird's-eye view of a building destroyed by a tornado.
A top-down view of a house damaged by a tornado.
A satellite image of a building destroyed by a tornado.
A bird's-eye view of a building damaged by a tornado.

Nepal Floods
An aerial view of houses surrounded by a flood.
A top-down view of houses damaged by floods.
A top-down view of a house damaged by floods inundated in water.
A satellite image of a building destroyed by a flood surrounded by water.
A satellite image of houses that was destroyed by a flood surrounded by water and trees.

Portugal Wildfire
An aerial view of forest land after it is torched by a wildfire.
An aerial view of buildings after a wildfire.
An aerial image of forest land scorched by a wildfire.
A bird's-eye view of a forest region with completely scorched trees.

Figure 4. Prompt Pool for xBD

3.4. Training using Synthetic Data

Following prior work in disaster assessment task [3], we adopt a siamese network with shared transformer backbone [11] for training. Specifically, we pass both pre and post images U and V using parameter-shared transformer backbones $\mathcal{E}_u$ and $\mathcal{E}_v$, resulting in feature embedding $f_u = \mathcal{E}_u(U)$ and $f_v = \mathcal{E}_v(V)$ respectively, each dimension d. We then fuse these embeddings by concatenating them to form $f \in \mathbb{R}^{2d}$, where $f = \text{concat}(f_s, f_d)$. Finally, we add a 2-layer MLP network $\mathcal{H}$ with a hidden dimension of d to predict a single output value indicating the probability of damaged building between the pre and post images. The whole network is then trained with a binary cross entropy loss using the binary ground truth labels.

However, directly training predictive models using only

synthetic data might result in poor accuracy due to the domain gap between synthetic and real images [38]. Therefore, we devise a two-stage training strategy to leverage the in-domain synthetic data, along with out-of-domain real data to effectively improve the target performance. In particular, we first train our network end-to-end including the encoders $\mathcal{E}_u$ and $\mathcal{E}_v$ as well as the MLP layers $\mathcal{H}$ using the source domain data $\mathcal{D}_s$. Subsequently, we follow prior work in robust learning [21] to fine-tune only the final layers of the MLP network $\mathcal{H}(.)$ using the synthetic data supervision from the target, while keeping the encoders fixed during the fine-tuning stage. We observed that only re-training the last layer prevents over-fitting the network to the synthetic data compared to complete end-to-end fine-tuning (Tab. 4), so we adopt this two-stage mechanism in our framework. During inference, we apply a sigmoid layer on top of the predicted output and threshold this probability to predict damaged buildings in the post-image if the predicted probability is > 0.5, and predict no damage otherwise.

Fine-tuning the MUSE model In the framework illustrated so far, we only use a frozen pre-trained MUSE model, where we fix the generative model itself and only use it for inference given input images and corresponding text prompts. However, such off-the-shelf models trained on billions of web-scale image-text data might contain images from a wide variety of domains, and might not be fully suited for use in specific domains like aerial or satellite imagery. Therefore, we also investigate the potential benefits offered by fine-tuning the pre-trained generative model for the specific task of aerial image classification. In particular, we collect the pre-disaster images from $\mathcal{D}_t$ and create prompts for each image from the *undamaged* pool to create a dataset of image-text pairs from the target domain. We then adopt adapter-fine tuning [18] to fine-tune the pre-trained model using these image-text pairs, which we found to be more resource-efficient than end-to-end fine-tuning. This fine-tuned model is expected to capture more domain specific properties unique to aerial and satellite imagery, and we compare this procedure with generation using the frozen model in Sec. 4.3.

4. Experiments

We next demonstrate the effectiveness of the proposed approach on several challenging transfer settings. We first introduce our choice of datasets in Sec. 4.1, specify the training details in Sec. 4.2 followed by the results in Sec. 4.3 and several ablations into our modeling and training choices in Sec. 4.4.

4.1. Datasets

xBD Dataset xBD [16] is a large-scale dataset designed for automatic disaster assessment using aerial and satellite imagery. The dataset covers synchronized pre- and post-event satellite imagery of both damaged and undamaged scenes from more than 19 events across the world, covering a variety of disaster types across varying severity levels. Since our focus in this paper is to improve robustness of aerial disaster assessment algorithms across disparate geographies, we choose 3 domains from xBD, namely *nepal-flooding*, *portugal-wildfires* and *moore-tornado* to demostrate our results, which have 36456, 18884 and 18491 images respectively. These domains encompass data from three distinct geographical subregions, each affected by entirely different types of disasters making it a challenging problem to improve cross-domain robustness.

SKAI Satellite Imagery In order to verify the effectiveness of our method in improving the performance across subtle domain variations, we adopt the SKAI dataset [23] consisting of pre and post hurricane images captured from different regions in the United States. The images in SKAI includes data collected from Ian, Maria, Michael and Laura hurricanes with 2733, 3709, 3991 and 3991 images respectively, which we use as the different domains for our cross-domain robustness setting. Note that both these datasets consist of heavy class imbalance, with more than 80% of the image-pairs capturing non-damaged buildings, adding an additional layer of complexity in bridging the domain shifts. For both the datasets, we show results using single-source and multi-source adaptation settings, in which we use supervised data from single source domain or all the domains except the target respectively.

4.2. Training and Evaluation Details

We use an Imagenet pretrained ViT-B/16 transformer backbone [42] as the encoder in our setting, and remove the last classification layer replacing it with the MLP head for binary classification. We then train the network using the two-stage approach discussed in Sec. 3.4, first using the supervised source domain images using Adam optimizer with a learning rate of 2e-6 for the pre-trained backbone and 2e-5 for the randomly initialized MLP layer, followed by re-training only the last MLP layer using synthesized target domain images using the same hyperparameters as above. We use a batch size of 64 in both cases and perform training for 5000 iterations. We use the validation images from the target domain to perform early stopping, which we observed to be very crucial to obtain good performance in our setting.

Following prior work in disaster assessment tasks from satellite imagery [16, 23], we adopt the AUPRC metric for evaluation which measures the area under the precision-recall curve across various thresholds, and is shown to be relatively more robust for cases like ours where there is severe class imbalance against positive examples. In terms of baselines, we compare with a source-only baseline which

Method	Ian →			Michael→			Laura→			Maria→			Avg.
	Michael	Laura	Maria	Ian	Laura	Maria	Ian	Michael	Maria	Ian	Michael	Laura	
Source Only	41.6	19.3	27.3	38.0	32.0	29.7	38.3	46.9	26.3	30.0	39.6	21.9	32.6
Ours w/ ZeroShot MUSE	49.2	**36.8**	**31.9**	47.4	**42.5**	**32.0**	**50.0**	54.7	**30.6**	42.6	**54.5**	**36.6**	**42.4** (+9.8%)
Ours w/ fine-tuned MUSE	**49.6**	29.9	25.8	**50.9**	31.5	28.8	49.2	**55.6**	26.4	**44.1**	53.6	27.6	39.4 (+6.8%)

Table 1. **Single-source Domain Adaptation Results on SKAI dataset** AUPRC values for different transfer settings from the SKAI dataset. We compare the results obtained by training using only real data from the source domain and combining it with synthetic generated data from the target domains on all the transfer settings. Evidently, our approach outperforms the source-only baseline setting new state-of-the-art.

Method	Moore-Tornado→		Nepal-Flooding→		Portugal-Wildfire→		Avg.
	Nepal-Flooding	Portugal-Wildfire	Moore-Tornado	Portugal-Wildfire	Moore-Tornado	Nepal-Flooding	
Source Only	23.8	23.2	14.5	18.5	45.3	24.7	25.0
Ours w/ ZeroShot MUSE	**49.5**	**24.1**	75.1	25.1	76.0	**51.6**	50.2 (+25.2%)
Ours w/ fine-tuned MUSE	43.9	**24.1**	**82.3**	**25.3**	**83.1**	47.5	**51.1** (+26.1%)

Table 2. **Single-source Domain Adaptation Results on xBD dataset.** AUPRC values for different transfer settings from the xBD dataset [16]. On each of the transfer setting, augmenting training using synthetic data from the target domain significantly outperforms the source-only baseline, with an improvement of 25.2% using a zeroshot generative model, and 26.1% with further fine-tuning the generative backbone on aerial image-text pairs.

only trains a predictive model on the source domain and evaluates on the target test-set. Since this does not use any target data, it serves as a fundamental baseline to illustrate the benefits obtained by our method. Note that prior UDA methods require both pre and post disaster images to learn domain agnostic features [4, 24, 46], preventing a direct comparison for our setting where only pre-disaster images from the target dataset are available.

4.3. Results

Single-source Zeroshot Adaptation We show the results for single-source UDA for domains from the SKAI dataset in Tab. 1 and xBD dataset in Tab. 2. As shown, our method of augmenting out of distribution training using synthetic images generated from MUSE model achieves better accuracy than the source-only baseline, with $\sim 10\%$ and $\sim 25\%$ improvements on the SKAI and the xBD datasets on average. Our improvements are consistent across all the transfer settings, with up to $\sim 70\%$ improvement on the more challenging cross-disaster cross-geography setting from xBD dataset, highlighting the effectiveness of leveraging generative foundational models to create synthetic data for low-resource domains even in expert tasks like disaster assessment.

Furthermore, we also compare the AUPRC results observed through fine-tuning the generative model on aerial imagery and satellite images, using the procedure outlined in Sec. 3.4. We observe that the model trained with data generated from fine-tuned model outperforms both the source-only baseline as well as the zeroshot settings on 4 out of 6 settings in xBD dataset with $\sim 1\%$ improvement on the average accuracy, indicating the potential in fine-tuning MUSE model on domain-specific images. On SKAI data however, we observe zeroshot model is better on the average AUPRC. A potential reason for this could be that the generative ability of the text-to-image generative model is reduced after fine-tuning on domain-specific images, impacting accuracy in few of the transfer settings, highlighting room for further improvement through more carefully designed fine-tuning strategies.

Multi-source Zeroshot Adaptation The comparison for both SKAI and xBD datasets on multi-source adaptation setting is shown in Tab. 3. Firstly, the accuracy achieved by multi-source models on all target domains is higher than single-source setting, which is expected since multi-source models have access to relatively more supervised data. Furthermore, the results from Tab. 3 clearly show the effectiveness of our approach even for such multi-source evaluation setting, where our method using zeroshot text-to-image generation yields 5.33% improvement over baseline on SKAI dataset and 29.13% improvement over baseline on the xBD dataset. Our benefits are consistent for both the datasets across all the transfer tasks, further supporting our hypothesis that text-to-image models can serve as strong data generators for low-resource domains. As seen for the case of single-source setting, we observe the gains yielded by data generation using zeroshot text-to-image models to be competitive when compared to fine-tuned models on both the datasets.

4.4. Ablations

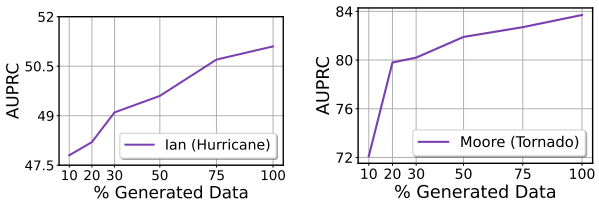
Ablations and Insights We show the effect of various design choices in our framework in Tab. 4. Firstly, we observe that training only using synthetic data without source domain data leads to poor results, potentially highlighting the limitations of synthetic data alone in training (`R1` vs `R4`). This facet of synthetic data has also been noted in prior works [38], indicating that manual supervision is still necessary to observe gains with synthetic supervision. Fur-

Dataset	SKAI-Dataset					xBD-Dataset			
Method	→Ian	→Michael	→Maria	→Laura	Avg.	→Moore-Tornado	→Nepal-Flooding	→Portugal-Wildfire	Avg.
Source Only	44.21	48.62	29.54	36.83	39.78	18.96	27.16	29.69	25.27
Ours w/ ZeroShot MUSE	**54.79**	52.24	**34.05**	**39.38**	**45.11** (+5.33%)	78.70	**52.40**	**32.10**	**54.40**(+29.13%)
Ours w/ fine-tuned MUSE	49.12	**53.83**	30.92	39.29	43.29 (+3.51%)	**83.18**	50.16	30.72	**54.69**(+29.42%)

Table 3. **Multi-source Domain Adaptation Results on SKAI and xBD datasets.** AUPRC values for different transfer settings, where we show the result of training using synthetic generated data from the respective target domain along with manual supervision from all the three remaining domains. Our approach outperforms the source-only baseline highlighting the effectiveness of training with generated synthetic data in bridging domain gaps.

Method	SKAI	xBD
(R0) Source Only	39.78	25.27
(R1) Only Synthetic Data	40.60	47.76
(R2) Joint Training on Real + Synthetic	43.11	49.66
(R3) End-to-end Finetuning	44.10	53.44
(R4) Last-Layer Finetune on SynData	**45.11**	**54.40**

Table 4. **Effect of training choices** We show the effect of various training choices in our framework, where last-layer re-training using only synthetic data (R4) outperforms training using only synthetic data without the source labels (R1), jointly training on both real and synthetic data (R2) as well as end-to-end finetuning using synthetic data (R3).



(a) Ian Hurricane (b) Moore Tornado

Figure 5. **Effect of the amount of generated synthetic data.** We show the positive influence of the volume of generated synthetic data (as a % of the target domain images) on the multi-source transfer setting, where adding more target data invariably helps to improve the final target accuracy for both SKAI (a) and xBD (b) datasets, with potential for further enhancement with more generated data.

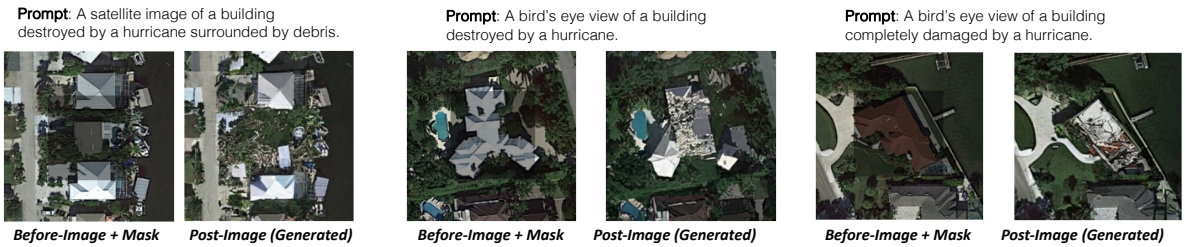
thermore, we also show that joint-training using real source and synthetic target datasets is inferior to our proposed approach of first pre-training on real source data followed by last-layer retraining on generated target domain data (R2 vs R4), supporting our two-stage training framework. Finally, we also observe that end-to-end fine-tuning using synthetic data under-performs the approach of finetuning the last layer only (R3 vs R4).

Effect of Volume of Synthetic Data We show the effect of the amount of generated synthetic data on the target AUPRC in Fig. 5. We observe that adding more synthetic data invariably helps the final target accuracy for both the datasets studied. More importantly, we observe no saturation even when using all target data to generate images indicating further room for improvement of target performance through low-cost synthetically generated data.

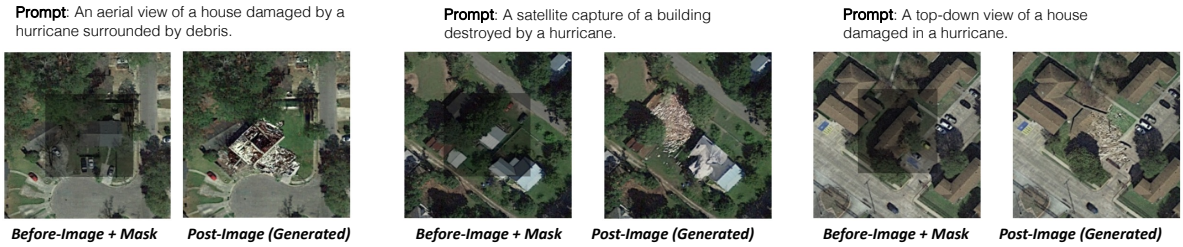
Visualizing Generated Images We show several illustrations of samples generated through our method in Fig. 6 on both SKAI (Fig. 6a, Fig. 6b) and xBD (Fig. 6c, Fig. 6d, Fig. 6e) datasets, where we include the pre-disaster image as well as the mask and the text-prompt used for conditional image editing through our generative model. In most cases, we observe that the text-to-image model incorporates the textual guidance and performs localized editing on the input image to generate a synthetic post-disaster image with great effectiveness. The model shows excellent capability in seamlessly handling the various types of disasters through our text-guidance, which helps to create realistic images in low-resource domains leading to significant empirical gains (Tab. 3).

5. Conclusion

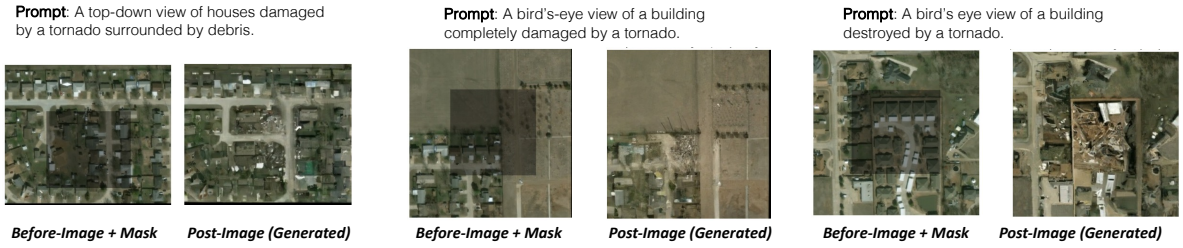
In this paper, we explore the potential of leveraging emerging text-to-image models in generating synthetic supervision to improve robustness across low-resource domains for disaster assessment tasks. We design an efficient and scalable data-generation pipeline by leveraging the localized image editing capabilities of transformer-based generative models [8]. Using this framework, we generate several thousand synthetic post-disaster images conditioned on pre-disaster images and text guidance, followed by a simple two-stage training mechanism that yields non-trivial benefits over a source-only baseline in both single source and multi-source domain adaptation setting. In terms of limitations, we noted a significant sensitivity of the training process to the quality and coherence of the generated synthetic data, which is directly affected by the presence of low-quality generated images. A potential future work can therefore be to additionally incorporate better filtering strategies into our framework to remove poor quality images and improve training. Nevertheless, our work serves as one of the first to explore the potential of text-to-image synthetic data for expert tasks like satellite disaster assessment, which holds massive potential



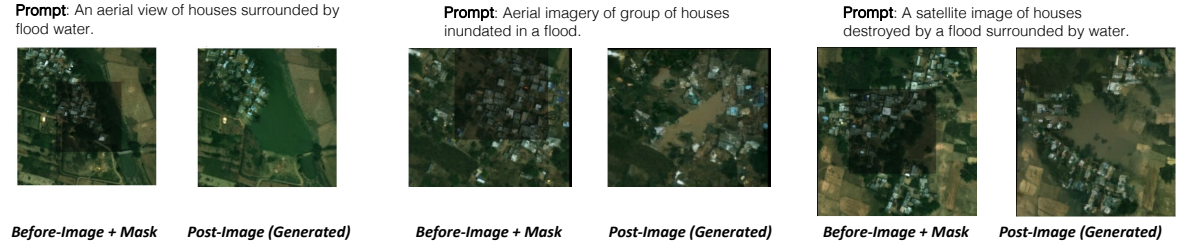
(a) Images from Ian domain.



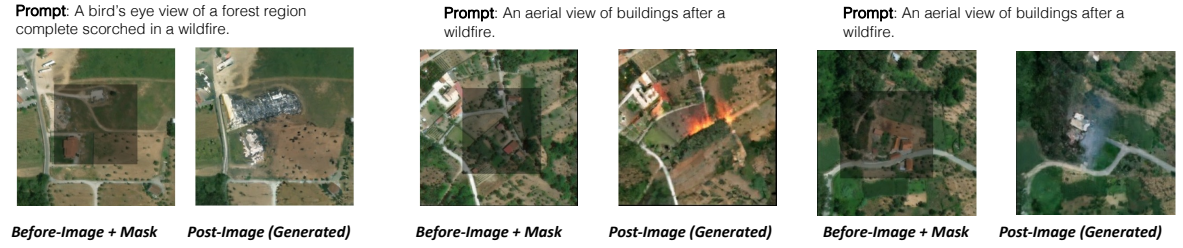
(b) Images from the Michael domain.



(c) Images from Moore-Tornado.



(d) Images from Nepal-Flooding.



(e) Images from Portugal-Wildfire.

Figure 6. **Visualization of text-to-image results.** We show several examples from our generated images, along with the pre-disaster image, correponding conditioning mask (overlapped on the pre-image) as well as the text-prompt used to generate the post-image from (a) Ian-Hurricane, (b) Michael-Hurricane, (c) Moore-Tornado, (d) Nepal-Floods and (e) Portugal-Wildfire.

for continued improvement with the development of better image generation models.

References

[1] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv*

preprint arXiv:2304.08466, 2023. 2

[2] Yanbing Bai, Junjie Hu, Jinhua Su, Xing Liu, Haoyu Liu, Xianwen He, Shengwang Meng, Erick Mas, and Shunichi Koshimura. Pyramid pooling module-based semi-siamese network: A benchmark model for assessing building damage from xbd satellite imagery datasets. *Remote Sensing*, 12(24): 4055, 2020. 2

[3] W. G. C. Bandara and Vishal M. Patel. A transformer-based siamese network for change detection. *IEEE International Geoscience and Remote Sensing Symposium*, 2022. 4

[4] Vitus Benson and Alexander Ecker. Assessing out-of-domain generalization for robust building damage detection, 2020. 1, 2, 3, 6

[5] Junchi Bin, Ran Zhang, Rui Wang, Yue Cao, Yufeng Zheng, Erik P. Blasch, and Zheng Liu. An efficient and uncertainty-aware decision support system for disaster response using aerial imagery. *Sensors (Basel, Switzerland)*, 22, 2022. 1

[6] Isabelle Bouchard, Marie-Ève Rancourt, Daniel Aloise, and Freddie Kalaitzis. On transfer learning for building damage assessment from satellite imagery in emergency contexts. *Remote Sensing*, 14(11), 2022. 1, 2

[7] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer. *CVPR*, 2022. 3

[8] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. Muse: Text-to-image generation via masked generative transformers, 2023. 1, 2, 3, 7

[9] Yifan Da, Zhiyuan Ji, and Yongsheng Zhou. Building damage assessment based on siamese hierarchical transformer framework. *Mathematics*, 2022. 2

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 3

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 4

[12] Xuefeng Du, Yiyou Sun, Jerry Zhu, and Yixuan Li. Dream the impossible: Outlier imagination with diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2

[13] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 12873–12883. Computer Vision Foundation / IEEE, 2021. 3

[14] Yunhao Ge, Jiashu Xu, Brian Nlong Zhao, Neel Joshi, Laurent Itti, and Vibhav Vineet. Beyond generation: Harnessing text to image models for object detection and segmentation. *arXiv preprint arXiv:2309.05956*, 2023. 2

[15] Rohit Gupta and Mubarak Shah. Rescuenet: Joint building segmentation and damage assessment from satellite imagery. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4405–4411. IEEE, 2021. 2

[16] Ritwik Gupta, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric Heim, Howie Choset, and Matthew Gaston. xbd: A dataset for assessing building damage from satellite imagery, 2019. 1, 2, 5, 6

[17] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 4

[18] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019. 5

[19] Tarun Kalluri, Astuti Sharma, and Manmohan Chandraker. Memsac: Memory augmented sample consistency for large scale domain adaptation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*, pages 550–568. Springer, 2022. 2

[20] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 2

[21] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022. 2, 5

[22] Xin Lai, Zhuotao Tian, Xiaogang Xu, Ying-Cong Chen, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. Decouplenet: Decoupled network for domain adaptive semantic segmentation. In *European Conference on Computer Vision*, 2022. 2

[23] Jihyeon Lee, Joseph Z. Xu, Kihyuk Sohn, Wenhan Lu, David Berthelot, Izzeddin Gur, Pranav Khaitan, Ke-Wei, Huang, Kyriacos Koupparis, and Bernhard Kowatsch. Assessing post-disaster damage from satellite imagery using semi-supervised learning techniques, 2020. 1, 2, 5

[24] Yundong Li, Chen Lin, Hongguang Li, Wei Hu, Han Dong, and Yi Liu. Unsupervised domain adaptation with self-attention for post-disaster building damage detection. *Neurocomputing*, 415:27–39, 2020. 1, 2, 3, 6

[25] Shaobo Lin, Kun Wang, Xingyu Zeng, and Rui Zhao. Explore the power of synthetic data on few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 638–647, 2023. 2

[26] Wen Lu, Lu Wei, and Minh Nguyen. Bitemporal attention transformer for building change detection and building damage assessment. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:4917–4935, 2024. 2

[27] Xiaopeng Lu, Lynnette Ng, Jared Fernandez, and Hao Zhu. Cigli: Conditional image generation from language & image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3134–3138, 2021. 2

[28] Hasan Nasrallah, Mustafa Shukor, and Ali J. Ghandour. Scinet: scale-invariant model for buildings segmentation from

aerial imagery. *Signal, Image and Video Processing*, 17:2999 – 3007, 2021. 1, 2

[29] Quang Nguyen, Truong Vu, Anh Tran, and Khoi Nguyen. Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[30] Abhishek V Potnis, Rajat C Shinde, Surya S Durbha, and Kuldeep R Kurte. Multi-class segmentation of urban floods from multispectral imagery using deep learning. In *IGARSS 2019-2019 IEEE international geoscience and remote sensing symposium*, pages 9741–9744. IEEE, 2019. 2

[31] Maryam Rahnemoonfar, Tashnim Chowdhury, Argho Sarkar, Debvrat Varshney, Masoud Yari, and Robin Roberson Murphy. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, 9:89644–89654, 2021. 2

[32] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 1, 2, 3

[33] Adam Roberts, Colin Raffel, Katherine Lee, Michael Matena, Noam Shazeer, Peter J Liu, Sharan Narang, Wei Li, and Yanqi Zhou. Exploring the limits of transfer learning with a unified text-to-text transformer. *Google, Tech. Rep.*, 2019. 3

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3

[35] 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*, pages 22500–22510, 2023. 2

[36] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2, 3

[37] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[38] Mert Bülent Sarıyıldız, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8011–8021, 2023. 1, 2, 5, 6

[39] Jie Shao, Ke Zhu, Hanxiao Zhang, and Jianxin Wu. Diffult: How to make diffusion model useful for long-tail recognition. *arXiv preprint arXiv:2403.05170*, 2024. 2

[40] Joonghyuk Shin, Minguk Kang, and Jaesik Park. Fill-up: Balancing long-tailed data with generative models. *arXiv preprint arXiv:2306.07200*, 2023. 2

[41] Wojciech Sirko, Sergii Kashubin, Marvin Ritter, Abigail Annkah, Yasser Salah Eddine Bouchareb, Yann Dauphin, Daniel Keysers, Maxim Neumann, Moustapha Cisse, and John Quinn. Continental-scale building detection from high resolution satellite imagery. *arXiv preprint arXiv:2107.12283*, 2021. 2

[42] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021. 5

[43] Yonglong Tian, Lijie Fan, Kaifeng Chen, Dina Katabi, Dilip Krishnan, and Phillip Isola. Learning vision from models rivals learning vision from data. *arXiv preprint arXiv: 2312.17742*, 2023. 2

[44] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2

[45] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7472–7481, 2018. 2

[46] Tinka Valentijn, Jacopo Margutti, Marc van den Homberg, and Jorma Laaksonen. Multi-hazard and spatial transferability of a cnn for automated building damage assessment. *Remote Sensing*, 12(17), 2020. 2, 6

[47] Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. Semantic image synthesis via diffusion models. *arXiv preprint arXiv: 2207.00050*, 2022. 2

[48] Xiang Wang, Yundong Li, Chen Lin, Yi Liu, and Shuo Geng. Building damage detection based on multi-source adversarial domain adaptation. *Journal of Applied Remote Sensing*, 15 (3):036503, 2021. 2

[49] Ethan Weber and Hassan Kané. Building disaster damage assessment in satellite imagery with multi-temporal fusion. *arXiv preprint arXiv:2004.05525*, 2020. 2

[50] Ethan Weber, Dim P. Papadopoulos, Àgata Lapedriza, Ferda Ofli, Muhammad Imran, and Antonio Torralba. Incidents1m: A large-scale dataset of images with natural disasters, damage, and incidents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:4768–4781, 2022. 1

[51] Chuyi Wu, Feng Zhang, Junshi Xia, Yichen Xu, Guoqing Li, Jibo Xie, Zhenhong Du, and Renyi Liu. Building damage detection using u-net with attention mechanism from pre-and post-disaster remote sensing datasets. *Remote Sensing*, 13(5): 905, 2021. 2

[52] Chuyi Wu, Feng Zhang, Junshi Xia, Yichen Xu, Guoqing Li, Jibo Xie, Zhenhong Du, and Ren yi Liu. Building damage detection using u-net with attention mechanism from pre- and post-disaster remote sensing datasets. *Remote. Sens.*, 13:905, 2021. 1

[53] Junshi Xia, Naoto Yokoya, Bruno Adriano, and Clifford Broni-Bediako. Openearthmap: A benchmark dataset for

global high-resolution land cover mapping. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6243–6253, 2022. 2

[54] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7452–7461, 2023. 2

[55] Joseph Z. Xu, Wenhan Lu, Zebo Li, Pranav Khaitan, and Valeriya Zaytseva. Building damage detection in satellite imagery using convolutional neural networks. *ArXiv*, abs/1910.06444, 2019. 2

[56] Tongkun Xu, Weihua Chen, Pichao Wang, Fan Wang, Hao Li, and Rong Jin. Cdtrans: Cross-domain transformer for unsupervised domain adaptation. *arXiv preprint arXiv:2109.06165*, 2021. 2

[57] Lihe Yang, Xiaogang Xu, Bingyi Kang, Yinghuan Shi, and Hengshuang Zhao. Freemask: Synthetic images with dense annotations make stronger segmentation models. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[58] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv: 2206.10789*, 2022. 2

[59] Zhuoran Yu, Chenchen Zhu, Sean Culatana, Raghuraman Krishnamoorthi, Fanyi Xiao, and Yong Jae Lee. Diversify, don't fine-tune: Scaling up visual recognition training with synthetic images. *arXiv preprint arXiv:2312.02253*, 2023. 2

[60] Tianjun Zhang, Yi Zhang, Vibhav Vineet, Neel Joshi, and Xin Wang. Controllable text-to-image generation with gpt-4. *arXiv preprint arXiv: 2305.18583*, 2023. 2

[61] Fei Zhao and Chengcui Zhang. Building damage evaluation from satellite imagery using deep learning. *2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 82–89, 2020. 1

[62] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. *Computer Vision and Pattern Recognition*, 2023. 2

[63] Yongchao Zhou, Hshmat Sahak, and Jimmy Ba. Training on thin air: Improve image classification with generated data. *arXiv preprint arXiv: 2305.15316*, 2023. 1, 2

[64] Jinjing Zhu, Haotian Bai, and Lin Wang. Patch-mix transformer for unsupervised domain adaptation: A game perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3561–3571, 2023. 2