Segment Any Change

Zhuo Zheng[1], **Yanfei Zhong**[2*], **Liangpei Zhang**[2], **Stefano Ermon**[1*]
[1]Stanford University
[2]Wuhan University
zhuozheng@cs.stanford.edu

Abstract

Visual foundation models have achieved remarkable results in zero-shot image classification and segmentation, but zero-shot change detection remains an open problem. In this paper, we propose the segment any change models (AnyChange), a new type of change detection model that supports zero-shot prediction and generalization on unseen change types and data distributions. AnyChange is built on the segment anything model (SAM) via our training-free adaptation method, bitemporal latent matching. By revealing and exploiting intra-image and inter-image semantic similarities in SAM's latent space, bitemporal latent matching endows SAM with zero-shot change detection capabilities in a training-free way. We also propose a point query mechanism to enable AnyChange's zero-shot object-centric change detection capability. We perform extensive experiments to confirm the effectiveness of AnyChange for zero-shot change detection. AnyChange sets a new record on the SECOND benchmark for unsupervised change detection, exceeding the previous SOTA by up to 4.4% F_1 score, and achieving comparable accuracy with negligible manual annotations (1 pixel per image) for supervised change detection. Code is available at https://github.com/Z-Zheng/pytorch-change-models.

1 Introduction

The Earth's surface undergoes constant changes over time due to natural processes and human activities. Some of the dynamic processes driving these changes (e.g., natural disasters, deforestation, and urbanization) have huge impact on climate, environment, and human life (Zhu et al., 2022). Capturing these global changes via remote sensing and machine learning is a crucial step in many sustainability disciplines (Yeh et al., 2020; Burke et al., 2021).

Deep change detection models have yielded impressive results via large-scale pre-training (Manas et al., 2021; Wang et al., 2022; Mall et al., 2023; Zheng et al., 2023) and architecture improvements (Chen et al., 2021b; Zheng et al., 2022). However, their capabilities depend on training data and are limited to specific application scenarios. These models cannot generalize to new change types and data distributions (e.g., new geographic areas) beyond those seen during training.

This desired level of generalization on unseen change types and data distributions requires change detection models with zero-shot prediction capabilities. However, the concept of *zero-shot change detection* has not been explored so far in the literature. While we are in the era of "foundation models" (Bommasani et al., 2021) and have witnessed the emergence of large language models (LLMs) and vision foundation models (VFMs) (e.g., CLIP (Radford et al., 2021) and Segment Anything Model (SAM) (Kirillov et al., 2023)) with strong zero-shot prediction and generalization capabilities via prompt engineering, zero-shot change detection is still an open problem.

To close this gap, we present *Segment Any Change*, the first change detection model with zero-shot generalization on unseen change types and data distributions. Our approach builds on SAM, the first

*Corresponding authors: Stefano Ermon, Yanfei Zhong

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-2581

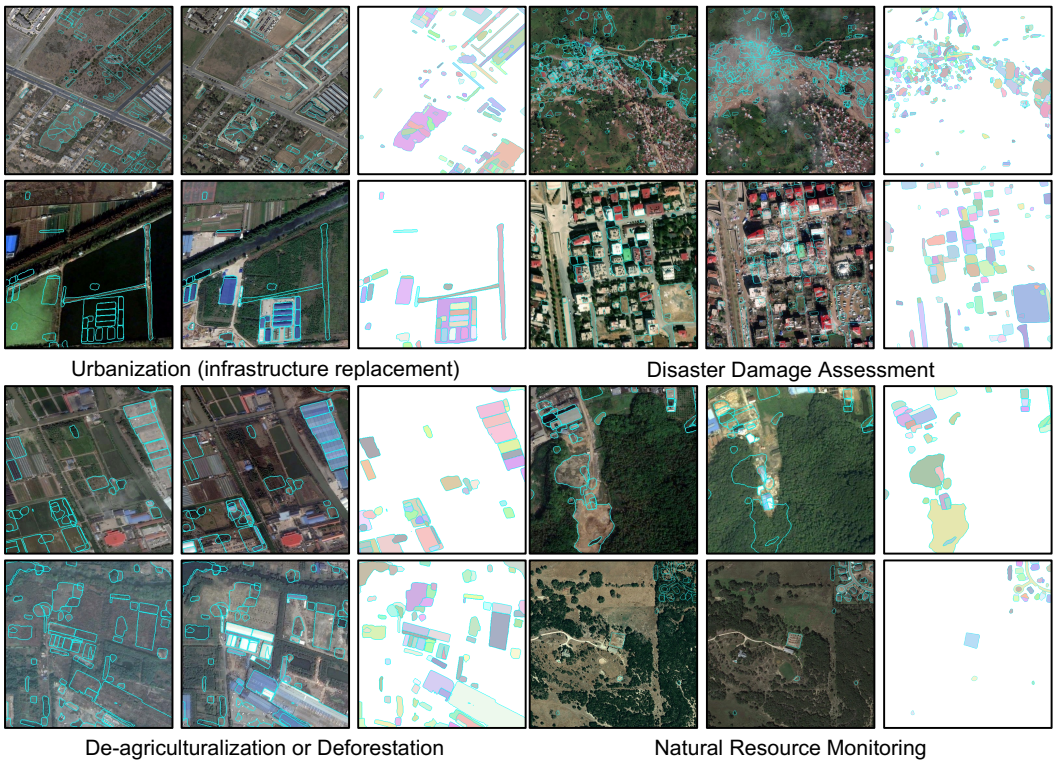


Figure 1: **Zero-Shot Change Detection with AnyChange** on a wide range of application scenarios in geoscience. Each subfigure presents the pre-event image, the post-event image, and their change instance masks in order. The boundary of each change instance mask is rendered by cyan, and meanwhile, these change masks are also drawn on pre/post-event images to show more clearly where the change occurred. The color of each change mask is used to distinguish between different instances.

promptable image segmentation model, which has shown extraordinary zero-shot generalization on object types and data distributions. While SAM is extremely capable, it is non-trivial to adapt SAM to change detection and maintain its zero-shot generalization and promptability due to the extreme data collection cost of large-scale change detection labels that would be required to enable promptable training as in the original SAM.

To resolve this problem, we propose a training-free method, *bitemporal latent matching*, that enables SAM to segment changes in bitemporal remote sensing images while inheriting these important properties of SAM (i.e., promptability, zero-shot generalization). This is achieved by leveraging intra-image and inter-image semantic similarities that we empirically discovered in the latent space of SAM when applying SAM's encoder on unseen multi-temporal remote sensing images. The resulting models, *AnyChange*, are capable of segmenting any semantic change.

AnyChange can yield class-agnostic change masks, however, in some real-world application scenarios, e.g., disaster damage assessment, there is a need for object-centric changes, e.g. to detect how many buildings are destroyed. To enable this capability we propose a point query mechanism for AnyChange, leveraging SAM's point prompt mechanism and our bitemporal latent matching for filtering desired object changes. The user only needs a single click on a desired object, AnyChange with the point query can yield change masks centered on this object's semantics, i.e., from this object class to others and vice versa, thus achieving object-centric change detection.

We demonstrate the zero-shot prediction capabilities of AnyChange on several change detection datasets, including LEVIR-CD (Chen & Shi, 2020), S2Looking (Shen et al., 2021), xView2 (Gupta et al., 2019), and SECOND (Yang et al., 2021). Due to the absence of published algorithms for zero-shot change detection, we also build baselines from the perspectives of zero-shot change proposal, zero-shot object-centric change detection, and unsupervised change detection. AnyChange outperforms other zero-shot baselines implemented by DINOv2 (Oquab et al., 2023) and SAM

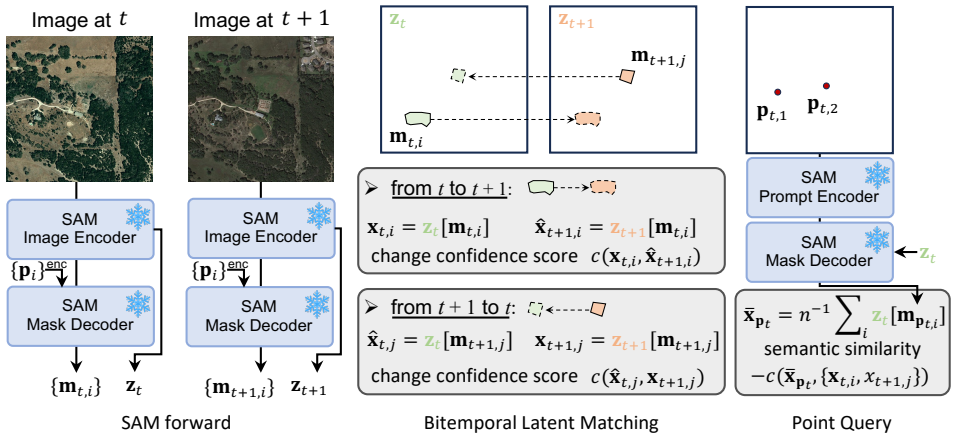


Figure 2: **Segment Any Change Models, AnyChange**. SAM forward: given grid points $\{\mathbf{p}_i\}$ as prompts and input images, SAM produces object masks $\{\mathbf{m}_{t,i}\}$ and image embedding $\mathbf{z_t}$ on the image at time t. **Bitemporal Latent Matching** does a bidirectional matching to compute the change confidence score for each change proposal, and then top-k sorting or thresholding is applied for zero-shot change proposal and detection. **Point Query** allows users to click some points (the case of two points in this subfigure) with the same category to filter class-agnostic change masks via semantic similarity for object-centric change detection.

with different matching methods in terms of zero-shot change proposal and detection. From the unsupervised change detection perspective, AnyChange beats the previous state-of-the-art model, I3PE (Chen et al., 2023), setting a new record of 48.2% F_1 on SECOND. We show some qualitative results in Fig. 1, demonstrating the zero-shot prediction capabilities of AnyChange on a wide range of application scenarios (i.e., urbanization, disaster damage assessment, de-agriculturalization, deforestation, and natural resource monitoring). The contributions of this paper are summarized as follows:

- **AnyChange**, the first zero-shot change detection model, enables us to obtain both instance-level and pixel-level change masks either in a fully automatic mode or interactively with simple clicks.
- **Bitemporal Latent Matching**, a training-free adaptation method, empowers SAM with zero-shot change detection by leveraging intra-image and inter-image semantic similarities of images in SAM's latent space.
- **Zero-Shot Change Detection** is explored for the first time. We demonstrate the effectiveness of AnyChange from four perspectives, i.e., zero-shot change proposal, zero-shot object-centric change detection, unsupervised change detection, and label-efficient supervised change detection, achieving better results over strong baselines or previous SOTA methods.

2 Related Work

Segment Anything Model (Kirillov et al., 2023) is the first foundation model for promptable image segmentation, possessing strong zero-shot generalization on unseen object types, data distributions, and tasks. The training objective of SAM is to minimize a class-agnostic segmentation loss given a series of geometric prompts. Based on compositions of geometric prompts, i.e., prompt engineering, SAM can generalize to unseen single-image tasks in a zero-shot way, including edge detection, object proposal, and instance segmentation. Our work extends SAM with zero-shot change detection for bitemporal remote sensing images via a training-free adaptation method, extending the use of SAM beyond single-image tasks.

Segment Anything Model for Change Detection. SAM has been used for change detection via a "parameter-efficient fine-tuning" (PEFT) paradigm (Mangrulkar et al., 2022), such as SAM-CD (Ding et al., 2023) that used Fast-SAM (Zhao et al., 2023) as a frozen visual encoder and fine-tuned adapter networks and the change decoder on change detection datasets in a fully supervised way. This model does not inherit the most two important properties of SAM, i.e., promptability and zero-shot generalization. Fine-tuning in a promptable way with large-scale training change data may achieve these two properties, however, collecting large-scale bitemporal image pairs with class-agnostic

https://doi.org/10.52202/079017-2581

change annotations is non-trivial (Tewkesbury et al., 2015; Zheng et al., 2023), thus no such method exists in the current literature. Our work introduces a new and economic adaptation method for SAM, i.e., training-free adaptation, guaranteeing these two properties with zero additional cost, making zero-shot change detection feasible for the first time.

Unsupervised Change Detection. The most similar task to zero-shot change detection is unsupervised change detection (Coppin & Bauer, 1996), however zero-shot change detection is a more challenging task. They both require models to find class-agnostic change regions, and the main difference is that zero-shot change detection also requires models to generalize to unseen data distributions. From early model-free change vector analysis (CVA) (Bruzzone & Prieto, 2000; Bovolo & Bruzzone, 2006) to advanced deep CVA (Saha et al., 2019) and I3PE (Chen et al., 2023), unsupervised change detection methods have undergone a revolution enabled by deep visual representation learning. These model-based unsupervised change detection methods need to re-train their models on new data distributions. Our proposed model is training-free and can achieve comparable or even better performance for unsupervised change detection.

3 Segment Any Change Models

This paper introduces *Segment Any Change* to resolve the long-standing open problem of zero-shot change detection (see Sec. 3.1). As illustrated in Fig. 2, we propose a new type of change detection models that support zero-shot prediction capabilities and generalization on unseen change types and data distributions, allowing two output structures (instance-level and pixel-level), and three application modes (fully automatic, semi-automatic with a custom threshold, and interactive with simple clicks). AnyChange achieves the above capabilities building on SAM in a training-free way.

3.1 Background

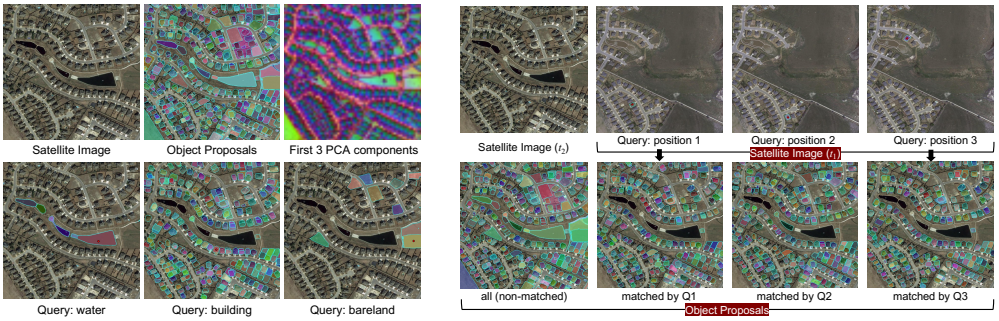
Preliminary: Segment Anything Model (SAM) is a promptable image segmentation model with an image encoder of ViT (Dosovitskiy et al., 2021), a prompt encoder, and a mask decoder based on transformer decoders. As Fig. 2(a) shows, given as input a single image at time t, the image embedding $\mathbf{z}_t$ is extracted from its image encoder. Object masks $\{\mathbf{m}_{t,i}\}$ can be obtained by its mask decoder given $\mathbf{z}_t$ and dense point prompts obtained by feeding grid points $\{\mathbf{p}_i\}$ into its prompt encoder.

Problem Formulation: Zero-Shot Change Detection is formulated at the pixel and instance levels. Our definition of "zero-shot" is similar to that of SAM, i.e., the model without training on change detection tasks can transfer zero-shot to change detection tasks and new image distributions. $\mathcal{U} = \{0, 1\}$ denotes a set of classes of non-change and change. A pixel position or instance area belongs to the change class if the corresponding semantic categories at different times are different. This means that the model should generalize to unseen change types, even though all change types are merged into a class of 1. The input is a bitemporal image pair $\mathbf{I}_t, \mathbf{I}_{t+1} \in \mathbb{R}^{h \times w \times *}$, where the image size is (h, w) and $*$ denotes the channel dimensionality of each image. For pixel level, the model is expected to yield a pixel-level change mask $\mathbf{C} \in \mathcal{U}^{h \times w}$. For instance level, the model is expected to yield an arbitrary-sized set of change masks $\{\mathbf{m}_i\}$, where each instance $\mathbf{m}_i$ is a polygon.

3.2 Exploring the Latent Space of SAM

Motivated by the experiments in Kirillov et al. (2023) on probing the latent space of SAM, we known there are potential semantic similarities between mask embeddings in the same natural image. We further explore the latent space for satellite imagery from both intra-image and inter-image perspectives, thus answering the following two questions:

Q1: Do semantic similarities exist on the same satellite image? Empirically, we find they do. We show this semantic similarity in two ways, i.e., visualizing the first components of principal components analysis (PCA) (Oquab et al., 2023) and probing the latent space (Kirillov et al., 2023), as Fig. 3 (a) shows. Observing the first three PCA components, geospatial objects with the same category have a similar appearance in this low-dimensional subspace. This suggests that this satellite image embedding from SAM encodes the semantics of geospatial objects reasonably well. Furthermore, we do the same latent space probing experiment as Kirillov et al. (2023) did, but on satellite images, and present the results in Fig. 3 (a) (bottom). We compute the mask embedding of object proposals and manually select three proposals with different categories (water, building, and bareland) as queries.



(a) intra-image semantic similarity (b) inter-image semantic similarity

Figure 3: Empirical evidence on semantic similarity on (a) the same satellite image and (b) the satellite images at different times. **(a)** The visualization of the first three PCA components indicates the objects with the same category are well matched with each other in SAM's latent space; Three red point queries confirm the existence of semantic similarity on the same satellite image. **(b)** The object proposals (indicated by red points) from the satellite image at t_1 as queries are used to match all proposals from the satellite image at t_2. (best viewed with zoom, especially for the point query) By matching the query embedding with other mask embeddings, we obtain the most similar object proposals with the query proposal. We find that the most similar object proposals mostly belong to the same category as the query.

Q2: Do semantic similarities exist on satellite images of the same location collected at different times? Empirically, we find they do. To verify this, we introduce a new satellite image from the same geographic area but at a different time t_1. The above satellite image is captured at time t_2. Different from the above latent space probing experiment, we use three object proposals with different spatial positions from the image at t_1, as queries. These three queries have the same category, i.e., building. By matching with all proposals from the image at t_2, we obtain three basically consistent results (F_1 of 68.1%±0.67% and recall of 96.2%±0.66%), as shown in Fig. 3 (b). This suggests that this semantic similarity exists on the satellite images at different times, even though the images have different imaging conditions because they were taken at different times.

From the above empirical study, we find that there are intra-image and inter-image semantic similarities in the latent space of SAM for unseen satellite images. These two properties are the foundation of our training-free adaptation method.

3.3 Bitemporal Latent Matching

Based on our findings, we propose a training-free adaptation method, namely *bitemporal latent matching*, which bridges the gap between SAM and remote sensing change detection without requiring for training or architecture modifications.

The main idea is to leverage the semantic similarities in latent space to identify changes in bitemporal satellite images. Given the image embedding $\mathbf{z}_t$ and object proposals $\mathcal{M}_t = \{\mathbf{m}_{t,i}\}_{i\in[1,2,\dots,N_t]}$ generated from SAM on the satellite image at time t, each object proposal $\mathbf{m}_{t,i} \in \mathbb{R}^{h\times w}$ is a binary mask. We can compute the mask embedding $\mathbf{x}_{t,i} = \mathbf{z}_t[\mathbf{m}_{t,i}] \in R^{d_m}$ by averaging the image embedding $\mathbf{z}_t$ over all non-zero positions indicated by the object proposal $\mathbf{m}_{t,i}$. Next, we introduce a similarity metric to measure the semantic similarity. To do this, we need to consider the statistical properties of the image embeddings from SAM. In particular, SAM's image encoder uses layer normalization (Ba et al., 2016). This means that the mask embedding has zero mean and unit variance if we drop the affine transformation in the last layer normalization, i.e., the variance $D(\mathbf{x}_{t,i}) = d_m^{-1}\sum_j(\mathbf{x}_{t,i}[j])^2 = 1$, thus we have the mask embedding's ℓ_2 norm $\|\mathbf{x}_{t,i}\|_2 = \sqrt{d_m}$, which is a constant since d_m is the channel dimensionality. Given this, cosine similarity is a suitable choice to measure similarity between two mask embeddings since they are on a hypersphere with a radius of $\sqrt{d_m}$, and differences are encoded by their directions. Therefore, we propose to use negative cosine similarity as the change confidence score $c(\mathbf{x}_i, \mathbf{x}_j)$ for mask embeddings $\mathbf{x}_i$ and $\mathbf{x}_j$:

$$c(\mathbf{x}_i, \mathbf{x}_j) = -\frac{\mathbf{x}_i \cdot \mathbf{x}_j}{d_m} \tag{1}$$

The next question is which two mask embeddings to use to compute the change confidence score. The real-world change is defined at the same geographic location from time t to $t+1$. This means

https://doi.org/10.52202/079017-2581

that it is comparable only if two mask embeddings cover the approximate same geographic region. Therefore, we additionally compute the mask embedding $\hat{\mathbf{x}}_{t+1,i} = \mathbf{z}_{t+1}[\mathbf{m}_{t,i}] \in \mathbb{R}^{d_m}$ on the image embedding $\mathbf{z}_{t+1}$ using the same object proposal $\mathbf{m}_{t,i}$. We then compute the change confidence score $c(\mathbf{x}_{t,i}, \hat{\mathbf{x}}_{t+1,i})$ for the change at $\mathbf{m}_{t,i}$ from t to $t+1$.

Since we need to guarantee the temporal symmetry (Zheng et al., 2021, 2022) of class-agnostic change, we propose to match the object proposals bidirectionally. To this end, the change confidence score $c(\mathbf{x}_{t+1,i}, \hat{\mathbf{x}}_{t,i})$ for the change at $\mathbf{m}_{t+1,i}$ from $t+1$ to t is also computed, where $\hat{\mathbf{x}}_{t,i} = \mathbf{z}_t[\mathbf{m}_{t+1,i}] \in \mathbb{R}^{d_m}$ is computed on the image embedding $\mathbf{z}_t$ with the same object proposal $\mathbf{m}_{t+1,i}$. Afterwards, we can match object proposals $\mathcal{M}_t$ and $\mathcal{M}_{t+1}$ bidirectionally, and $(N_t + N_{t+1})$ change proposals with their confidence score are obtained in total. We propose to finally obtain change detection predictions by sorting by confidence scores and selecting the top-k elements or by angle thresholding. The pseudo-code of Bitemporal Latent Matching is in Appendix B.

3.4 Point Query Mechanism

To empower AnyChange with interactive change detection with semantics, we combine our bitemporal latent matching with the point prompt mechanism of SAM, thus yielding the point query mechanism. Given as input a set of single-temporal point $\{\mathbf{p}_{t,i}\} = (x_{t,i}, y_{t,i})$ with the same category, t denote that this point belongs to the image at time t, and (x, y) indicates the spatial coordinate of image domain. The object proposals $\{\mathbf{m}_{\mathbf{p}_{t,i}}\}$ can be obtained via SAM's point prompts. Following our bitemporal latent matching, we then compute their average mask embedding $\bar{\mathbf{x}}_{\mathbf{p}_t} = n^{-1} \sum_1^n \mathbf{x}_{\mathbf{p}_{t,i}}$ and match it with all proposals $\{\mathcal{M}_t, \mathcal{M}_{t+1}\}$ via cosine similarity. In this way, the object-centric change detection results can be obtained via a custom angle threshold.

4 Experiments

In this section, we demonstrate the two most basic applications of AnyChange, i.e., (i) zero-shot change proposal and detection and (ii) change data engine. We conduct experiments from these two perspectives to evaluate our method.

4.1 Zero-Shot Object Change Proposals

Datasets: We use four commonly used change detection datasets to evaluate AnyChange. The first three, i.e., LEVIR-CD (Chen & Shi, 2020), S2Looking (Shen et al., 2021), and xView2 (Gupta et al., 2019), are building-centric change detection datasets. SECOND (Yang et al., 2021) is a multi-class (up to 36 change types) urban change detection dataset with full annotation. For zero-shot object proposal evaluation, we convert their labels into binary if the dataset has multi-class change labels.
Metrics: Conventional change detection mainly focuses on pixel-level evaluation using F_1, precision, and recall. More and more real-world applications have started to focus on instance-level change, also called "change parcel". Based on this requirement and AnyChange's capability of predicting instance change, we adapt the evaluation protocol of the zero-shot object proposal (Kirillov et al., 2023) for the zero-shot change proposal since they have the same output structure. The metric is mask AR@1000 (Lin et al., 2014). Note that change proposals are class-agnostic, therefore, for the first three building-centric datasets, we cannot obtain accurate F_1 and precision due to incomplete "any change" annotations. These two metrics are only used for reference and to see whether the model is close to the naive baseline (predict all as "change" class, and vice versa). Here we mainly focus on recall for both pixel- and instance-levels.
Baselines: AnyChange is based on SAM, however, there is no SAM or other VFM[2]-based zero-shot change detection model that can be used for comparison in the current literature. For a fair comparison, we build three strong baselines based on DINOv2 (Oquab et al., 2023) (a state-of-the-art VFM) or SAM. The simple baseline is CVA (Bruzzone & Prieto, 2000), a model-free unsupervised change detection method based on ℓ_2 norm as dissimilarity and thresholding. We build "DINOv2+CVA", an improved version with DINOv2 using an idea similar to DCVA (Saha et al., 2019). We build "SAM+Mask Match", which follows the macro framework of AnyChange and replaces the latent match with the mask match that adopts the IoU of masks as the similarity. "SAM+CVA Match" follows the same idea of AnyChange but adopts the negative ℓ_2 norm of feature difference as the similarity to compute pixel-level change map via SAM feature-based CVA first. The instance-level voting is then adopted to obtain change proposals. We also build each "Oracle" version of AnyChange as an upper bound, where we fine-tune SAM with LoRA ($r = 32$) (Hu et al., 2022) and train a change

[2]Visual foundation model, like SAM, CLIP, and DINOv2, etc.

Table 1: **Zero-shot Object Change Proposals**. The metrics include pixel-based F_1, Precision (Prec.), and Recall (Rec.) and instance-based mask AR@1000. Note that the metric names rendered with gray represent inaccurate estimations due to the absence of ground truth of "any change", but reflect whether their predictions approximate the naive baseline (predict all as "change" class).

		LEVIR-CD				**S2Looking** (binary)				**xView2** (binary)				**SECOND** (binary)			
Method	Backbone	F_1	Prec.	Rec.	mask AR↑	F_1	Prec.	Rec.	mask AR↑	F_1	Prec.	Rec.	mask AR↑	F_1 ↑	Prec.	Rec.	mask AR↑
• *pixel level*																	
CVA	-	12.2	7.5	32.6	-	5.8	3.1	44.3	-	7.6	4.3	33.3	-	30.2	26.5	35.2	-
DINOv2+CVA	ViT-G/14	17.3	9.5	96.6	-	4.3	2.2	92.9	-	5.9	3.1	62.0	-	41.4	26.9	89.4	-
◦ *pixel and instance level*																	
SAM+Mask Match	ViT-B	12.2	8.7	20.2	6.8	4.7	2.6	28.6	15.1	8.6	8.7	23.5	10.2	23.5	30.3	19.2	7.2
SAM+CVA Match	ViT-B	12.7	7.5	41.9	9.0	3.7	1.9	78.3	31.5	3.0	1.6	29.3	12.2	34.1	23.9	59.5	18.6
AnyChange	ViT-B	23.4	13.7	83.0	**32.6**	7.4	3.9	94.0	**48.3**	13.4	7.6	59.3	**27.8**	**44.6**	30.5	83.2	**27.0**
AnyChange (Oracle)	ViT-B	73.3	65.2	83.6	37.3	60.3	53.4	69.1	29.8	49.7	38.7	69.5	31.6	69.5	68.2	70.8	16.8
SAM+Mask Match	ViT-L	16.8	11.9	28.8	13.4	4.0	2.3	13.6	8.5	6.8	4.3	15.7	7.1	16.0	29.0	11.0	4.2
SAM+CVA Match	ViT-L	13.2	7.2	88.2	29.7	3.0	1.5	95.5	40.3	2.6	1.3	26.6	11.6	35.3	22.0	90.1	25.5
AnyChange	ViT-L	21.9	12.5	87.1	**39.5**	6.5	3.3	93.0	**50.1**	9.8	5.3	66.1	**30.5**	**42.3**	27.9	87.4	**28.6**
AnyChange (Oracle)	ViT-L	75.3	65.4	88.7	44.8	62.2	57.6	67.6	30.6	51.4	37.5	81.7	38.8	69.1	63.0	76.5	19.9
SAM+Mask Match	ViT-H	17.8	12.6	30.2	16.1	4.1	2.5	13.1	8.6	5.5	3.5	13.3	6.3	14.2	28.8	9.5	3.7
SAM+CVA Match	ViT-H	13.2	7.1	92.3	36.8	2.9	1.5	96.3	41.2	3.1	1.6	34.1	15.7	35.6	22.1	91.1	25.7
AnyChange	ViT-H	23.0	13.3	85.0	**43.4**	6.4	3.3	93.2	**50.4**	9.4	5.1	62.2	**29.3**	**41.8**	27.4	88.7	**29.0**
AnyChange (Oracle)	ViT-H	76.3	64.7	93.1	51.4	61.0	53.2	71.3	31.6	50.8	36.4	84.2	40.4	69.5	63.7	76.5	19.3

confidence score network on each dataset. More implementation details can be seen in Appendix C.1.

Results: We compare the change recall of AnyChange with other zero-shot baselines in Table 1. For *pixel-level change recall*, AnyChange achieves better recalls than the other two SAM baselines, especially when using a small ViT-B. This is because bitemporal latent matching better measures semantic similarity, i.e., using the angle between two embeddings. We can observe that the gap between AnyChange and SAM baselines gradually reduces as the backbone becomes larger since visual representation capabilities generally become stronger. AnyChange still has better average performance on four datasets, although a stronger representation can close the performance gap to some extent. This highlights the importance of finding the essential semantic difference. Besides, our AnyChange with ViT-H (636M parameters) achieves comparable recalls to CVA with DINOv2 (ViT-G, 1,100M parameters) on four datasets in fewer parameters. On the SECOND dataset, all variants of AnyChange achieve better zero-shot change detection performance than the strong baseline, DINOv2+CVA, and the margin is up to 3.2% F_1. For *instance-level change recall*, AnyChange outperforms the other two SAM baselines by a significant margin. This further confirms the effectiveness and superiority of bitemporal latent matching. We observe that the Oracles obtained via supervised learning have superior precision since they learn semantics and dataset bias explicitly, however, this comes with the cost of recall on pixel or instance levels.

Ablation: Matching Strategy. "Mask Match" performs geometry-based matching, while "CVA Match" and bitemporal latent matching perform latent-based matching. In Table 1, we see that latent-based matching is much more promising than geometry-based matching from multiple perspectives of pixel-level and instance-level change recall and zero-shot change detection performance. Compared with "CVA Match", AnyChange outperforms "SAM+CVA Match" on instance-level object change proposals by large margins and has comparable recalls on pixel-level change recalls. In principle, the similarity of "CVA Match" is linearly related to our bitemporal latent matching since the magnitude of the embeddings is a constant. Therefore, the performance difference lies in the computational unit (pixel or instance). "CVA Match" is pixel-wise, while bitemporal latent matching is instance-wise. This suggests that it is more robust to compute mask embedding averaged over all involved pixel embeddings for matching.

Ablation: Matching Direction. As presented in Table 2, we can find that the performance of single-directional matching is sensitive to temporal order, e.g., mask AR of two single-directional matching on LEVIR-CD are 1.3% and 35.9%, respectively. This is because the class-agnostic change is naturally temporal symmetric which is exactly the motivation of our bidirectional design. This result also confirms generally higher and more robust zero-shot change proposal capability.

Ablation: Robustness to radiation variation. We used ViT-B as the backbone for fast experiments. The results are presented in Table 3. The performance jitter of mask AR is less than 2% (-1.9%,+0.1%,

https://doi.org/10.52202/079017-2581

Table 2: **Ablation: Matching Direction**. The backbones are ViT-B. Single-directional matching is sensitive to temporal order, while bidirectional matching is more stable due to its guaranteed temporal symmetry.

	LEVIR-CD				**S2Looking** (binary)				**xView2** (binary)				**SECOND** (binary)			
Direction	F_1	Prec.	Rec.	mask AR↑	F_1	Prec.	Rec.	mask AR↑	F_1	Prec.	Rec.	mask AR↑	F_1 ↑	Prec.	Rec.	mask AR↑
bidirectional	23.4	13.7	83.0	32.6	7.4	3.9	94.0	48.3	13.4	7.6	59.3	27.8	44.6	30.5	83.2	27.0
from t to $t+1$	17.7	10.1	72.8	1.3	9.0	4.8	85.6	32.1	15.3	9.0	49.8	27.3	41.2	31.2	60.6	14.8
from $t+1$ to t	23.6	13.6	88.7	35.9	8.1	4.3	79.5	19.4	12.3	7.6	32.7	6.7	46.1	34.1	71.3	14.9

Table 3: **Ablation: Robustness to radiation variation**. The backbones are ViT-B. Radiation variation was simulated by applying random color jitter independently to the pre- and post-event images.

	LEVIR-CD				**S2Looking** (binary)				**xView2** (binary)				**SECOND** (binary)			
Condition	F_1	Prec.	Rec.	mask AR↑	F_1	Prec.	Rec.	mask AR↑	F_1	Prec.	Rec.	mask AR↑	F_1 ↑	Prec.	Rec.	mask AR↑
baseline	23.4	13.7	83.0	32.6	7.4	3.9	94.0	48.3	13.4	7.6	59.3	27.8	44.6	30.5	83.2	27.0
w/ color jitter	22.6	13.1	84.2	30.7	7.4	3.8	94.1	48.4	13.5	7.7	54.7	25.9	42.4	28.7	81.7	26.4

-1.9%, -1.4%) on these four datasets. We believe this sensitivity to radiation variation is acceptable for most applications.

4.2 Zero-shot Object-centric Change Detection

Zero-shot object change proposal yields class-agnostic change masks. The point query mechanism can convert the class-agnostic change into object-centric change via simple clicks of the user on a single-temporal image, thus providing an interactive mode for AnyChange. This step is typically easy for humans. Here we evaluate the point query on three building-centric change datasets.

Results: We demonstrate the effect of the point query in Fig. 4. Without the point query, AnyChange yields class-agnostic change masks, including building changes, vegetation changes, etc. With a single-point query on a building, we can observe that change masks unrelated to the building are filtered out. Further clicking two more buildings to improve the stability of mask embedding, we find the building changes previously missed are successfully recovered. Table 4 quantitatively reflects this mechanism. With a single-point query, the zero-shot performances on three datasets significantly gain $\sim$ 15% F_1 score. This improvement hurts recall as a cost, however it achieves a better trade-off between precision and recall.

After increasing to three-point queries, the recalls of the model on three datasets get back to some extent, and the model has comparable precision with the single-point query. The zero-shot performances on three datasets gain $\sim$ 3% F_1 further. These results confirm the effectiveness of the point query as a plugin for AnyChange to provide an interactive mode.

Table 4: **Zero-shot Object-centric Change Detection.** All results of introducing semantics from the point query are accurate estimations since the detected changes are object-centric.

	LEVIR-CD			**S2Looking** (binary)			**xView2**(binary)		
Method	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.
AnyChange-H	23.0	13.3	85.0	6.4	3.3	93.2	9.4	5.1	62.2
+ Point Query									
1 point query	38.6	30.9	51.4	21.8	16.9	34.1	25.1	24.8	25.4
	↑15.6			↑15.4			↑15.7		
3 point queries	42.2	28.5	81.1	24.0	14.5	68.5	28.1	20.0	47.2
	↑19.2			↑17.6			↑18.7		

4.3 AnyChange as Change Data Engine

AnyChange can provide pseudo-labels for unlabeled bitemporal image pairs with zero or few annotation costs. To evaluate this capability, we conduct experiments on two typical tracks for remote sensing change detection: supervised object change detection and unsupervised change detection.

Training recipe: On S2Looking, we train the model on its training set with pure pseudo-labels of AnyChange with ViT-H with a single-point query. All training details follow Zheng et al. (2023) except the loss function. On SECOND, we train the model on its training set with pure pseudo-labels of AnyChange with ViT-H via fully automatic mode, and other details follow Zheng et al. (2022). For

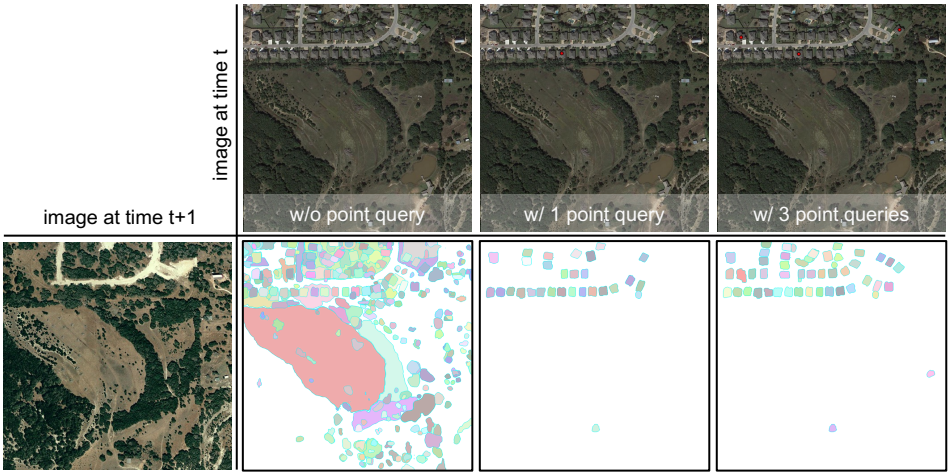


Figure 4: Examples of **Point Query Mechanism**. The effects of w/o point query, one-point query, and three-point queries are shown in sequence from left to right. (best viewed digitally with zoom, especially for the red points)

Table 5: **Supervised Object Change Detection**. Comparison with the state-of-the-art change detectors on the **S2Looking** `test`. "R-18": ResNet-18. The amount of Flops was computed with a float32 tensor of shape [2,512,512,3] as input.

Method	Backbone	Fine-tuning on	#labeled pixels↓	F_1 ↑	Prec.	Rec.	#Params.	#Flops
FC-Siam-Diff (Daudt et al., 2018)	-	100% GT	1.2×10^{10}	13.1	83.2	15.7	1.3M	18.7G
STANet (Chen & Shi, 2020)	R-18	100% GT	1.2×10^{10}	45.9	38.7	56.4	16.9M	156.7G
CDNet (Chen et al., 2021a)	R-18	100% GT	1.2×10^{10}	60.5	67.4	54.9	14.3M	-
BiT (Chen et al., 2021b)	R-18	100% GT	1.2×10^{10}	61.8	72.6	53.8	11.9M	34.7G
ChangeStar (1×96) (Zheng et al., 2021)	R-18	100% GT	1.2×10^{10}	66.3	70.9	62.2	16.4M	65.3G
+ `Changen-90k` (Zheng et al., 2023)	R-18	100% GT	1.2×10^{10}	67.1	70.1	64.3	16.4M	65.3G
ChangeStar (1×96) (Zheng et al., 2021)	MiT-B1	100% GT	1.2×10^{10}	64.3	69.3	59.9	18.4M	67.3G
+ `Changen-90k` (Zheng et al., 2023)	MiT-B1	100% GT	1.2×10^{10}	67.9	70.3	65.7	18.4M	67.3G
ChangeStar (1×96)	R-18	1% GT	1.4×10^{8}	37.2	63.1	26.3	16.4M	65.3G
ChangeStar (1×96)	R-18	0.1% GT	4.7×10^{6}	9.2	10.0	8.5	16.4M	65.3G
ChangeStar (1×96) (Ours)	R-18	AnyChange	3.5×10^{3}	40.2	40.4	39.9	16.4M	65.3G

both these two cases, the loss function is BCE loss with a label smoothing of 0.4. See Appendix C.2 for more details.

Supervised Object Change Detection Results. Table 5 presents standard benchmark results on the S2Looking dataset, which is the one of most challenging building change detection datasets. From a data-centric perspective, we use the same architecture with different fine-tuned data. The model architecture adopts ResNet18-based ChangeStar (1×96) (Zheng et al., 2021) due to its simplicity and good performance. We set the fine-tuning on 100% ground truth as the upper bound, which achieved 66.3% F_1. We can observe that fin-tuning on the pseudo-labels of AnyChange with ViT-H yields 40.2% F_1 with 3,500-pixel annotations[3]. Leveraging AnyChange, we achieve 61% of the upper bound at a negligible cost ($\sim 10^{-5}$% full annotations). We also compare it to the model trained with fewer annotations (1% and 0.1%). We find that their performances are reduced by a significant amount, and are inferior to the model trained with pseudo-labels from AnyChange. This confirms the potential of AnyChange as a change data engine for supervised object change detection.

Unsupervised Change Detection Results. AnyChange's class-agnostic change masks are natural pseudo-labels for unsupervised change detection. We also compare our AnyChange with unsupervised change detection methods. In Table 6, we find that AnyChange with ViT-B in a zero-shot setting improves over the previous state-of-the-art method, I3PE (Chen et al., 2023). To learn the biases on the SECOND dataset, we trained a ChangeStar (1×256) model with pseudo-labels of AnyChange on the SECOND training set. The setting follows I3PE

[3]The training set of S2Looking has 3,500 image pairs. We apply AnyChange with a single-point query for each pair to produce pseudo-labels. Therefore, the number of manual annotations is 3,500 pixels.

https://doi.org/10.52202/079017-2581

(Chen et al., 2023), thus we align their backbone and use ResNet50 for ChangeStar (1×256). The results show that two variants of unsupervised ChangeStar (1×256) outperform I3PE. Notably, based on pseudo-labels of AnyChange with ViT-B, our model set a new record of 48.2% F_1 on the SECOND dataset for unsupervised change detection. Besides, we obtain some useful insights for unsupervised change detection: (*i*) deep features from VFMs significantly assist unsupervised change detection models since DINOv2+CVA beats all advanced competitors except I3PE. Before our strong baseline, DINOv2+CVA, CVA has been always regarded as a simple and ineffective baseline for unsupervised change detection. (*ii*) dataset biases are helpful for in-domain unsupervised change detection since our model trained with pseudo-labels achieves higher performance than these pseudo-labels. This indicates the model learns some biases on the SECOND dataset, which may be change types and style.

Table 6: **Unsupervised Change Detection**. Comparison with the state-of-the-art unsupervised change detectors on the **SECOND** `test`. "*" indicates this change detection model is trained with pseudo labels predicted by AnyChange.

Method	Backbone	F_1 ↑	Prec.	Rec.
ISFA (Wu et al., 2013)	-	32.9	29.8	36.8
DSFA (Du et al., 2019)	-	33.0	24.2	51.9
DCAE (Bergamasco et al., 2022)	-	33.4	35.7	31.4
OBCD (Xiao et al., 2016)	-	34.3	29.6	40.7
IRMAD (Nielsen, 2007)	-	34.5	28.6	43.6
KPCA-MNet (Wu et al., 2021)	-	36.7	29.5	48.5
DCVA (Saha et al., 2019)	-	36.8	29.6	48.7
DINOv2+CVA (zero-shot, our impl.)	ViT-G/14	41.4	26.9	**89.4**
I3PE (Chen et al., 2023)	R-50	43.8	**36.3**	55.3
AnyChange (zero-shot, ours)	ViT-H	41.8	27.4	88.7
+ ChangeStar (1×256)* (ours)	R-50	45.0	30.2	88.2
AnyChange (zero-shot, ours)	ViT-B	44.6	30.5	83.2
+ ChangeStar (1×256)* (ours)	R-50	**48.2**	33.5	86.4

5 Conclusion

We present the segment any change models (AnyChange), a new type of change detection model for zero-shot change detection, allowing fully automatic, semi-automatic with custom threshold, and interactive mode with simple clicks. The foundation of all these capabilities is the intra-image and inter-image semantic similarities in SAM's latent space we identified on multi-temporal remote sensing images. Apart from zero-shot change detection, we also demonstrated the potential of AnyChange as the change data engine and demonstrated its superiority in unsupervised and supervised change detection. AnyChange is an out-of-the-box zero-shot change detection model, and a step forward towards a "foundation model" for the Earth vision community.

Acknowledgements

This work was supported in part by ARO (W911NF-21-1-0125), ONR (N00014-23-1-2159), the CZ Biohub, and the National Natural Science Foundation of China under Grant No. 42325105.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Luca Bergamasco, Sudipan Saha, Francesca Bovolo, and Lorenzo Bruzzone. Unsupervised change detection using convolutional-autoencoder multiresolution features. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19, 2022.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Francesca Bovolo and Lorenzo Bruzzone. A theoretical framework for unsupervised change detection based on change vector analysis in the polar domain. *IEEE Transactions on Geoscience and Remote Sensing*, 45(1): 218–236, 2006.

Lorenzo Bruzzone and Diego F Prieto. Automatic analysis of the difference image for unsupervised change detection. *IEEE Transactions on Geoscience and Remote sensing*, 38(3):1171–1182, 2000.

Marshall Burke, Anne Driscoll, David B Lobell, and Stefano Ermon. Using satellite imagery to understand and promote sustainable development. *Science*, 371(6535):eabe8628, 2021.

Hao Chen and Zhenwei Shi. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sensing*, 12(10):1662, 2020.

Hao Chen, Wenyuan Li, and Zhenwei Shi. Adversarial instance augmentation for building change detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2021a.

Hao Chen, Zipeng Qi, and Zhenwei Shi. Remote sensing image change detection with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2021b.

Hongruixuan Chen, Jian Song, Chen Wu, Bo Du, and Naoto Yokoya. Exchange means change: An unsupervised single-temporal change detection framework based on intra-and inter-image patch exchange. *ISPRS Journal of Photogrammetry and Remote Sensing*, 206:87–105, 2023.

Pol R Coppin and Marvin E Bauer. Digital change detection in forest ecosystems with remote sensing imagery. *Remote sensing reviews*, 13(3-4):207–234, 1996.

Rodrigo Caye Daudt, Bertr Le Saux, and Alexandre Boulch. Fully convolutional siamese networks for change detection. In *ICIP*, pp. 4063–4067. IEEE, 2018.

Lei Ding, Kun Zhu, Daifeng Peng, Hao Tang, and Haitao Guo. Adapting segment anything model for change detection in hr remote sensing images. *arXiv preprint arXiv:2309.01429*, 2023.

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. In *ICLR*, 2021.

Bo Du, Lixiang Ru, Chen Wu, and Liangpei Zhang. Unsupervised deep slow feature analysis for change detection in multi-temporal remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(12):9976–9992, 2019.

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. *arXiv preprint arXiv:1911.09296*, 2019.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment Anything. In *ICCV*, pp. 4015–4026, October 2023.

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 *ECCV*, pp. 740–755. Springer, 2014.

Utkarsh Mall, Bharath Hariharan, and Kavita Bala. Change-aware sampling and contrastive learning for satellite images. In *CVPR*, pp. 5261–5270, 2023.

Oscar Manas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *ICCV*, pp. 9414–9423, 2021.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2022.

Allan Aasbjerg Nielsen. The regularized iteratively reweighted mad method for change detection in multi-and hyperspectral data. *IEEE Transactions on Image processing*, 16(2):463–478, 2007.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021.

 https://doi.org/10.52202/079017-2581

Sudipan Saha, Francesca Bovolo, and Lorenzo Bruzzone. Unsupervised deep change vector analysis for multiple-change detection in vhr images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(6):3677–3693, 2019.

Li Shen, Yao Lu, Hao Chen, Hao Wei, Donghai Xie, Jiabao Yue, Rui Chen, Shouye Lv, and Bitao Jiang. S2looking: A satellite side-looking dataset for building change detection. *Remote Sensing*, 13(24):5094, 2021.

Andrew P Tewkesbury, Alexis J Comber, Nicholas J Tate, Alistair Lamb, and Peter F Fisher. A critical synthesis of remotely sensed optical image change detection techniques. *Remote Sensing of Environment*, 160:1–14, 2015.

Di Wang, Jing Zhang, Bo Du, Gui-Song Xia, and Dacheng Tao. An empirical study of remote sensing pretraining. *IEEE Transactions on Geoscience and Remote Sensing*, 2022.

Chen Wu, Bo Du, and Liangpei Zhang. Slow feature analysis for change detection in multispectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 52(5):2858–2874, 2013.

Chen Wu, Hongruixuan Chen, Bo Du, and Liangpei Zhang. Unsupervised change detection in multitemporal vhr images based on deep kernel pca convolutional mapping network. *IEEE Transactions on Cybernetics*, 52 (11):12084–12098, 2021.

Pengfeng Xiao, Xueliang Zhang, Dongguang Wang, Min Yuan, Xuezhi Feng, and Maggi Kelly. Change detection of built-up land: A framework of combining pixel-based detection and object-based recognition. *ISPRS Journal of Photogrammetry and Remote Sensing*, 119:402–414, 2016.

Kunping Yang, Gui-Song Xia, Zicheng Liu, Bo Du, Wen Yang, Marcello Pelillo, and Liangpei Zhang. Asymmetric siamese networks for semantic change detection in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–18, 2021.

Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature communications*, 11(1):2583, 2020.

Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything. *arXiv preprint arXiv:2306.12156*, 2023.

Zhuo Zheng, Ailong Ma, Liangpei Zhang, and Yanfei Zhong. Change is Everywhere: Single-temporal supervised object change detection in remote sensing imagery. In *ICCV*, pp. 15193–15202, 2021.

Zhuo Zheng, Yanfei Zhong, Shiqi Tian, Ailong Ma, and Liangpei Zhang. ChangeMask: Deep multi-task encoder-transformer-decoder architecture for semantic change detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 183:228–239, 2022.

Zhuo Zheng, Shiqi Tian, Ailong Ma, Liangpei Zhang, and Yanfei Zhong. Scalable multi-temporal remote sensing change data generation via simulating stochastic change process. In *ICCV*, pp. 21818–21827, 2023.

Zhe Zhu, Shi Qiu, and Su Ye. Remote sensing of land change: A multifaceted perspective. *Remote Sensing of Environment*, 282:113266, 2022.

A Appendix / supplemental material

B Pseudocode of Bitemporal Latent Matching

Algorithm 1 Bitemporal Latent Matching (top-k)

Require: $\mathbf{z}_t, \mathbf{z}_{t+1}$: image embeddings; $\mathcal{M}_t, \mathcal{M}_{t+1}$: object proposals; k: number of change proposals

for $i = 1$ **to** N_t **do**
  $\mathbf{m}_{t,i} \leftarrow \mathcal{M}_t[i]$
  $\mathbf{x}_{t,i} \leftarrow \mathbf{z}_t[\mathbf{m}_{t,i}]$
  $\hat{\mathbf{x}}_{t+1,i} \leftarrow \mathbf{z}_{t+1}[\mathbf{m}_{t,i}]$
  $c_{t,i} \leftarrow -d_m^{-1}\mathbf{x}_{t,i} \cdot \hat{\mathbf{x}}_{t+1,i}$ ▷ compute change confidence score via embedding dissimilarity
for $j = 1$ **to** N_{t+1} **do**
  $\mathbf{m}_{t+1,j} \leftarrow \mathcal{M}_{t+1}[j]$
  $\mathbf{x}_{t+1,j} \leftarrow \mathbf{z}_{t+1}[\mathbf{m}_{t+1,j}]$
  $\hat{\mathbf{x}}_{t,j} \leftarrow \mathbf{z}_t[\mathbf{m}_{t+1,j}]$
  $c_{t+1,j} \leftarrow -d_m^{-1}\mathbf{x}_{t+1,j} \cdot \hat{\mathbf{x}}_{t,j}$ ▷ compute change confidence score via embedding dissimilarity
$\{\mathbf{m}_i\}_1^k \leftarrow$ `sort`$(\mathcal{M}_t \cup \mathcal{M}_{t+1}$, `key=lambda t,i:`$\{c_{t,i}\} \cup \{c_{t+1,j}\}$`[t,i])[:`$k$`]` ▷ Python-style top-k sorting
return $\{\mathbf{m}_i\}_1^k$

C Implementation Details

C.1 Baselines for zero-shot change proposal and detection

SAM forward: For object proposal generation, we adopt a point per side of 64, an NMS threshold of 0.7, a predicted IoU threshold of 0.5, and a stability score threshold is 0.8 for LEVIR-CD, S2Looking, SECOND, and 0.95 for xView2. To obtain the image embedding with a constant ℓ_2 norm ($\sqrt{d_m}$), we demodulate the output of the image encoder f of SAM with the affine transformation $(\mathbf{w}; \mathbf{b})$ of the last layer normalization, i.e., $\mathbf{z} = \mathbf{w}^{-1}(f(\mathbf{x}) - \mathbf{b})$, where $\mathbf{x}$ is the single input image. The forward computation was conducted on 8 NVIDIA A4000 GPUs.

DINOv2 + CVA: The implementation is straightforward where the image embedding is first extracted from DINOv2 and then upsampled into the image size with bilinear interpolation. The change intensity map is computed by the ℓ_2 norm of the difference between bitemporal image embeddings. The optimal threshold is obtained by an adaptive threshold selection method, OTSU (Otsu, 1979). We also conduct a linear search for its optimal threshold on a small validation set, and the searched results have comparable performance with OTSU. Considering optimal peak performance and ease of use, we choose OTSU's threshold as the default for this baseline.

SAM + Mask Match: The main idea of this baseline is to use the geometric difference to measure an object's change. In general, if an object disappears at the next time image, the object mask generated by SAM is empty or has a different geometric shape, and vice versa. Therefore, using their geometric shape difference to measure if the change occurred is reasonable. To this end, we compute pairwise mask IoU between pre-event object masks and post-event object masks to match a potentially identical object at another time for each object. We recognize the object region belongs to non-change when successfully matched (IoU$>$0.5), otherwise this region belongs to a change.

SAM + CVA Match: This baseline follows the same idea of AnyChange, i.e., latent-based matching, but adopts the negative ℓ_2 norm of feature difference as the similarity. This method first computes pixel-level change map via SAM feature-based CVA, the procedure of which is similar to "DINOv2 + CVA". The instance-level voting with a threshold of 0.5 is then adopted to obtain change proposals, which means that each region is considered as a change when more than half of the pixels are identified as changes.

AnyChange: For a fair comparison and automatic benchmark evaluation, we use fully automatic mode for AnyChange in Table 1. The change angle threshold (155°) is obtained by OTSU and a linear search on the small validation set sampled from the SECOND training set. This threshold is directly used for the other three datasets without any adjustment.

https://doi.org/10.52202/079017-2581

AnyChange (Oracle): We only train a LoRA ($r = 32, alpha = 320, dropout = 0.1$) for the SAM model on each dataset due to unaffordable full-parameter training. Besides, we attach a change confidence score network on the image encoder of SAM to predict an accurate semantic similarity instead of our negative cosine similarity. This network architecture is composed of an MLP block (Linear-LayerNorm-GELU-Linear-LayerNorm), an upsampling block (ConvTranspose-LayerNorm-GELU-ConvTranspose-Linear-LayerNorm-Linear-LayerNorm) for 4× upsampling, and a linear layer for predicting change confidence score. The loss function is a compound of binary cross-entropy loss and soft dice loss. The inference pipeline exactly follows AnyChange, where it first generates object masks and computes mask embeddings, and the change confidence score is obtained from the trained score network. The training iterations are 200 epochs with a batch size of 16, AdamW optimizer with a weight decay of 0.01. The learning rate schedule is "poly" ($\gamma = 0.9$) decay with an initial learning rate of 6e-5 The training data augmentation adopts random rotation, flip, scale jitter, and cropping. The crop size is 512 for LEVIR-CD and S2Looking and 256 for xView2 and SECOND, respectively.

C.2 Pseudo-label Training

We have the training step of 20k, a batch size of 16, SGD with a momentum of 0.9, and a weight decay of 1e-3 for optimization on the SECOND training set with pseudo-labels of AnyChange in fully automatic mode. For regularization, we adopt the label smoothing of 0.4 and strong data augmentations which include random crop to 256×256, flip, rotation, scale jitter, and temporal-wise adjustment of brightness and contrast.

D Limitations

Zero-shot change detection is an open problem in remote sensing and computer vision communities. There is little reference to point out any potential effective roadmap before our work. Our work is the first to define this problem suitably and provide a simple yet effective model and evaluation protocol. The problem formulation and evaluation protocol themselves bring some potential limitations (i.e., scenario coverage, the robustness to objects of different geometries) since there is little mature infrastructure, e.g., a concept-complete class-agnostic change detection dataset.

E Broader Impacts

The proposed method can detect class-agnostic changes in Earth's surface, however, its actual effect is impacted by SAM's latent space. The model may produce some impossible changes due to SAM's biases. These issues warrant further research and consideration when building upon this work for real-world applications.

F More visualization

Nature Image Domain. Our method can be also used in the natural image domain to detect more general object changes, as shown in Fig. 5.

More demonstration of point query mechanism. We provide an additional example to supplement Fig. 4, which includes unchanged and changed buildings simultaneously. It is more clear to show that AnyChange can more accurately detect building changes with the help of the point query.

Effectiveness of AnyChange in detecting tiny object changes. Fig. 7 demonstrates a case of tiny/minor changes, such as small vehicle changes. We observe that directly applying AnyChange to the original image overlooks these subtle changes (see the first row). After we bilinearly upsampled the red box region by 2× and then applied AnyChange to it, we find some tiny/minor changes could be detected (see the second row). This observation shows that our method has the ability on tiny object change detection, although it is not yet optimal. Future work could use our approach as a strong baseline to further improve the detection of subtle changes.

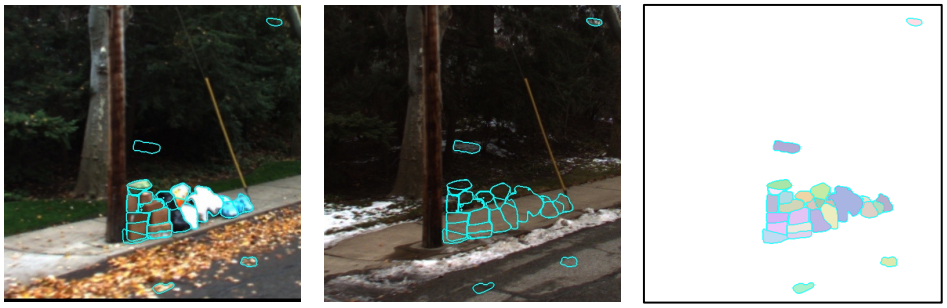


Figure 5: AnyChange on **Natural Image Domain**. Best viewed digitally with zoom.

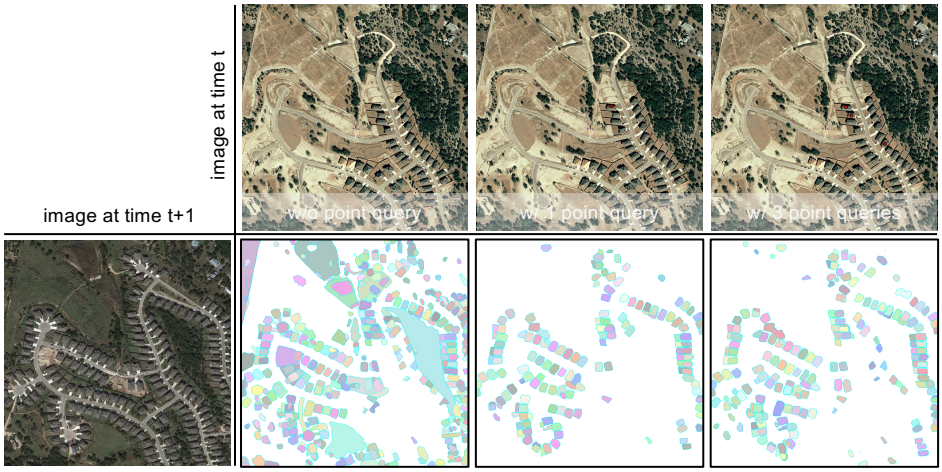


Figure 6: Examples of **Point Query Mechanism**. The effects of w/o point query, one-point query, and three-point queries are shown in sequence from left to right. (best viewed digitally with zoom, especially for the red points)

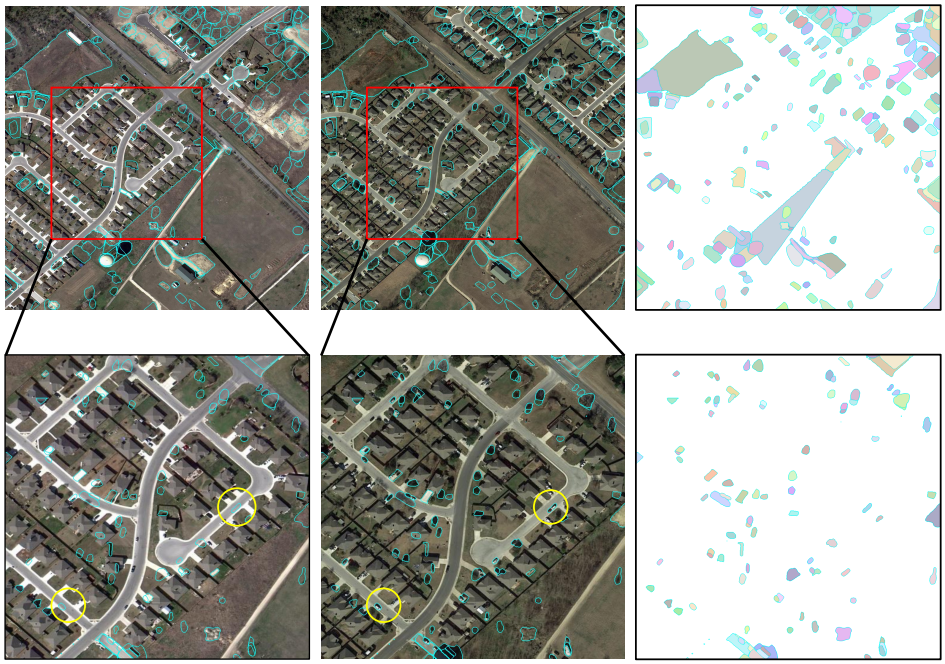


Figure 7: Illustration of the effectiveness of AnyChange in detecting tiny or minor object changes.

https://doi.org/10.52202/079017-2581

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have made claims in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the limitations in Appendix D

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: We have no theoretical result.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have fully described technical details and pseudo-code for our method.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

https://doi.org/10.52202/079017-2581

Answer: [Yes]

Justification: We provide source code and pseudo-code for our method.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We have reported all training and test details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Our algorithm is training-free and deterministic.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have provided this information.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: conformed.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have discussed broader impacts in Appendix E

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

https://doi.org/10.52202/079017-2581

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper poses no such risks

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have cited their publications.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: Our paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-2581