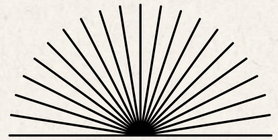
CV PROJECT PRESENTATION

Enhancing Change Detection in “The Change You Want to See”

TEAM ID:
19

PRESENTED BY:
Amey Choudhary (2021113017)
Hardik Mittal (2021114016)

PRESENTED TO:
Makarand Tapaswi
Anoop M. Namboodiri



Overview of the Project

01 Given two images of the same scene, being able to **automatically detect the changes** in them has practical applications in a variety of domains. In this paper, the authors tackle the change detection problem with the goal of **detecting "object-level" changes in an image pair despite differences in their viewpoint and illumination** (photometric and geometric changes)

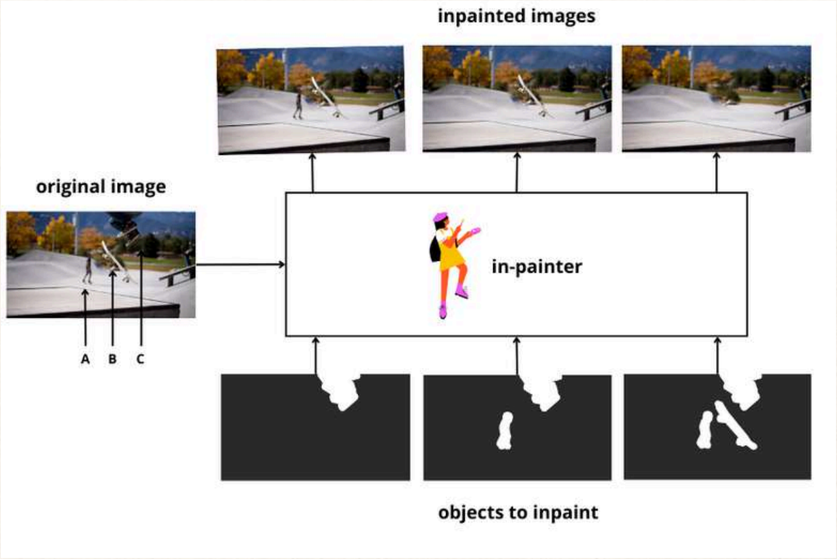
02 The paper makes the following four contributions:
(i) Propose a scalable methodology for **obtaining a large-scale change detection** training dataset by leveraging existing object segmentation benchmarks
(ii) Introduce **a co-attention based novel architecture** that is able to implicitly determine correspondences between an image pair and find changes in the form of bounding box predictions
(iii) Contribute **four evaluation datasets** that cover a variety of domains and transformations, including synthetic image changes, real surveillance images of a 3D scene, and synthetic 3D scenes with camera motion
(iv) **Evaluate the model on these four datasets** and demonstrate zero-shot and beyond training transformation generalization.

Figure 1. In this image pair, 5 out of 6 differences are shown using yellow boxes. Can you spot the remaining one? Our model can.

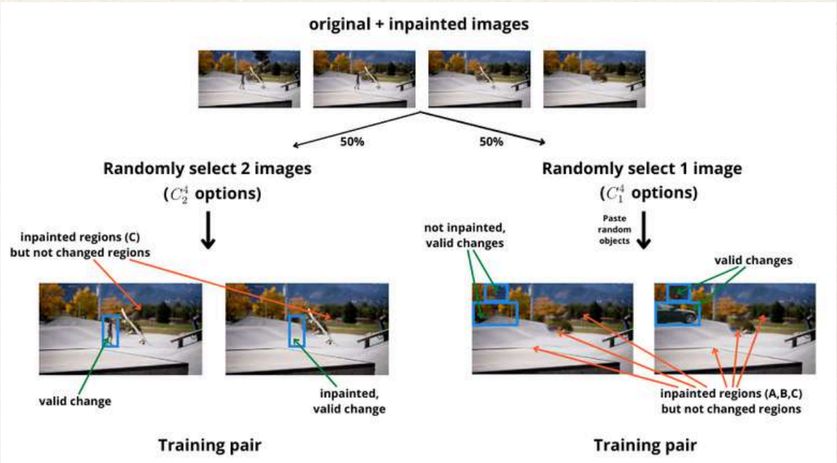
Obtaining change detection datasets

01 For Training

Name of the dataset	Number of Images	Type of Transformations (done to the dataset)
Coco-inpainted Using the CoCo dataset, which already contains the bounding box, LaMa (an inpainting tool) was used to make different object disappear in multiple locations in the same image along with adding random objects into different version	57,000 + 3,000	Random Affine (scale, translation, rotation) Photometric (colour jittering)



In-painting method to compute several images with inpainted regions



Randomly sample an image pair for training, along with their ground truth bounding boxes

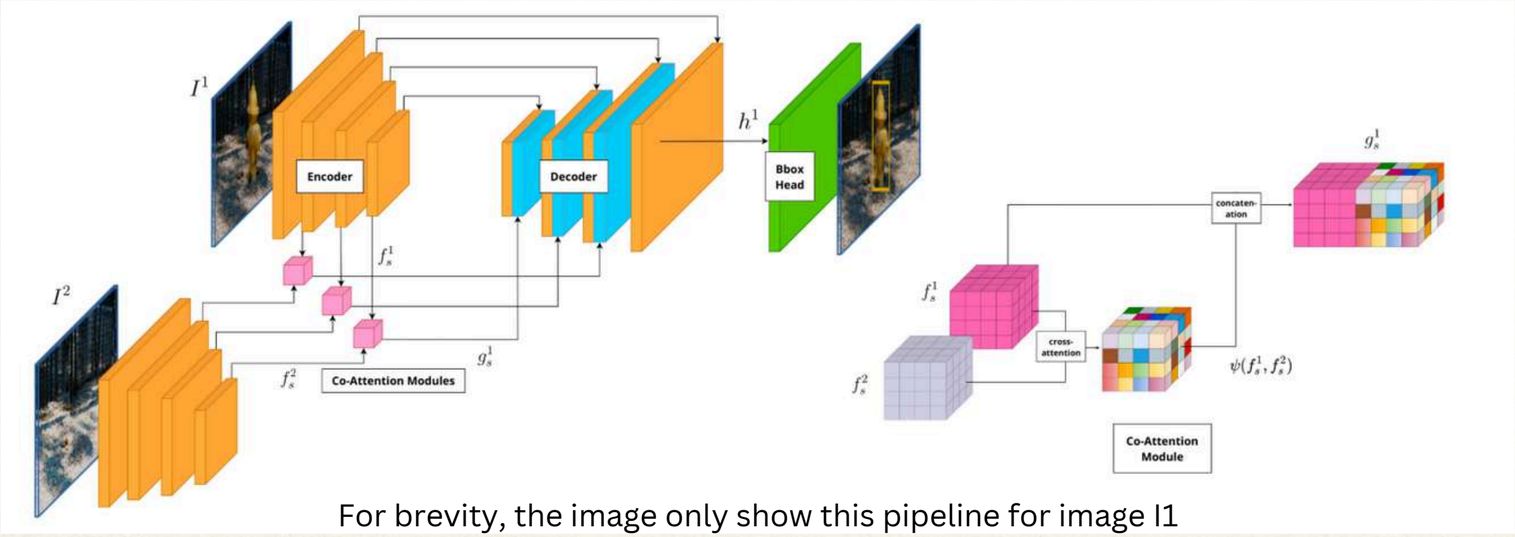
Obtaining change detection datasets

02 For Testing

Name of the dataset	Number of Images	Type of Transformations (done to the dataset)
Coco-inpainted (small, medium, large)	1655 + 1747 + 1006	Random Affine (scale, translation, rotation) Photometric (colour jittering)
Synthtext-Change	5000	Random text to background images (No photometric or geometric changes)
VIRAT-STD (Created bouding boxes for ground truth)	1000	Temporal Changes based on CCTV No photometric or geometric changes
Kubric-Change (Photo-realistic Simulation)	1605	3D Camera center moves (3D geometric changes)

*Sample images examples in the appendix

Architecture



For brevity, the image only show this pipeline for image I1

Given two images I1, I2, an **UNet encoder (using CNNs)** produces feature maps f1s , f2s respectively at multiple resolutions. A **coattention module** is then used to compute conditioned feature maps g1s (concatenation of I1 + attention of I1+I2) , g2s (vice versa) that are impicitly registered with the other image.

A **U-Net style decoder** (with skip connections modulated with scSE blocks) is then applied to the original and conditioned features maps to produce feature maps h1, h2.

Finally, the **centernet based bbox detector head** uses h1, h2 to produce bounding boxes for I1, I2 respectively.

Objectives and Goals in proposal

Since the paper is 2 years old, we aimed to reproduce and enhance the results of the paper using the current best techniques available today

Goal # 1

Integration of Advanced Backbone Architectures & Variations

Goal # 2

Try to integrate PINNs based loss in the architecture

Goal # 3

Enhanced Attention Mechanisms

Reproducing the results claimed

01 Final Results (ResNet50, 3 COAM layers, with scSE blocks)

Dataset Name	Claims (mAP)	Our results (mAP)
COCO-Inpainted	0.63	0.5964
Synthtext-Change	0.89	0.8836
VIRAT-STD	0.54	0.54092
Kubric-Change	0.76	0.66386

We deduce that with each run, random transformations were happening for all the datasets which led to different results.

We had checked it by running the same config multiple times and we were getting different results. Thus the authors would have probably chosen the best results among multiple runs

Reproducing the results claimed

02 Other results (the scores are the mAP scores for the configurations)

Backbone	# attn modules	attn type	scSE	geom transformations	Claims	Our results
ResNet18	2	COAM	✗	affine	0.11	0.28
ResNet18	3	COAM	✗	affine	0.37	0.376
ResNet50	3	NOAM	✗	affine	0.21	0.185
ResNet50	3	COAM	✗	affine	0.58	0.433
ResNet50	3	COAM	✓	affine	0.63	0.596
ResNet50	3	COAM	✓	identity	0.73	0.739
ResNet50	3	NOAM	✓	identity	0.79	0.804

Shifting to 100 epochs

- Since 200 epochs was taking 4 days to train, we asked Makarand and he suggested to shift to 100 epochs

Dataset Name	200 epochs (mAP)	100 epochs (mAP)
COCO-Inpainted	0.5964	0.5278
Synthtext-Change	0.8836	0.8535
VIRAT-STD	0.54092	0.5241
Kubric-Change	0.66386	0.6583

As we can see that there is no considerable change in shifting from 200 to 100 epochs, therefore we didrest of our experiments on 100 epochs only

Our novelties

- **Improved Backbone :** We shifted the backbone from ResNet18/50 to using a **Mix ViT** and **ResNext50** with the hypothesis that since both the models perform better than ResNet50 on ImageNet object detection task they should be performing better than these models on our similar task as well

- **PINNs inspired loss :**
  - **Consistency Loss:**
    - The consistency loss penalizes inconsistencies between the predictions for the left and right images.
    - It computes the mean squared difference between the final predictions of the left and right images.
    - By minimizing this loss, the model is encouraged to make consistent predictions for corresponding regions in both images.
  - **Spatial Continuity Loss:**
    - The spatial continuity loss encourages spatial continuity in the predicted changes.
    - It uses the total variation loss, which penalizes abrupt changes in the predictions.
    - The total variation loss is computed separately for the left and right predictions and then summed.
    - By minimizing this loss, the model learns to produce spatially smooth and coherent predictions, avoiding discontinuities or isolated regions.

- **Change in the type of attention :** Added multi-head attention and relative position embeddings, making it more expressive and capable of capturing long-range dependencies
  - By incorporating relative position embeddings, the attention mechanism can capture both content-based relationships (through the dot product between query and key vectors) and positional relationships (through the relative position embeddings). This allows the model to better understand the spatial structure and relationships within the feature maps

Our novelties

02 Other results

Configuration	attn type	Coco-Inpainted	Synthtext-Change	VIRAT-STD	Kubric-Change
ResNet50	COAM	0.5964	0.8836	0.5409	0.6639
ResNext50	COAM	0.5767	0.8763	0.5451	0.6952
PINN losses	COAM	0.3929	0.6686	0.1999	0.2671
New Attention Mechanism	COAM	0.20996	0.16279	0.04217	0.02321
ResNet50	NOAM	0.1849	0.0231	0.0201	0.009
Mix Vision Transformer	NOAM	0.3375	0.0818	0.0081	0.0081

Link to WandB runs for reproduction and novelties: https://wandb.ai/wb-team-hardik/cyws

Challenged Faced

- There were multiple dependancy issues which had to be solved and took some time (we already successfully merged one pull request into the author’s repository to fix one of them)

- The main issue that held us back from trying more new ideas and more hyperparameter tuning was
  - The huge amount of training time (4 days) for one config
  - Lack of availability of 4 2080Ti gpus in Ada at once for which we had to wait for days sometimes

Thank you

Appendix

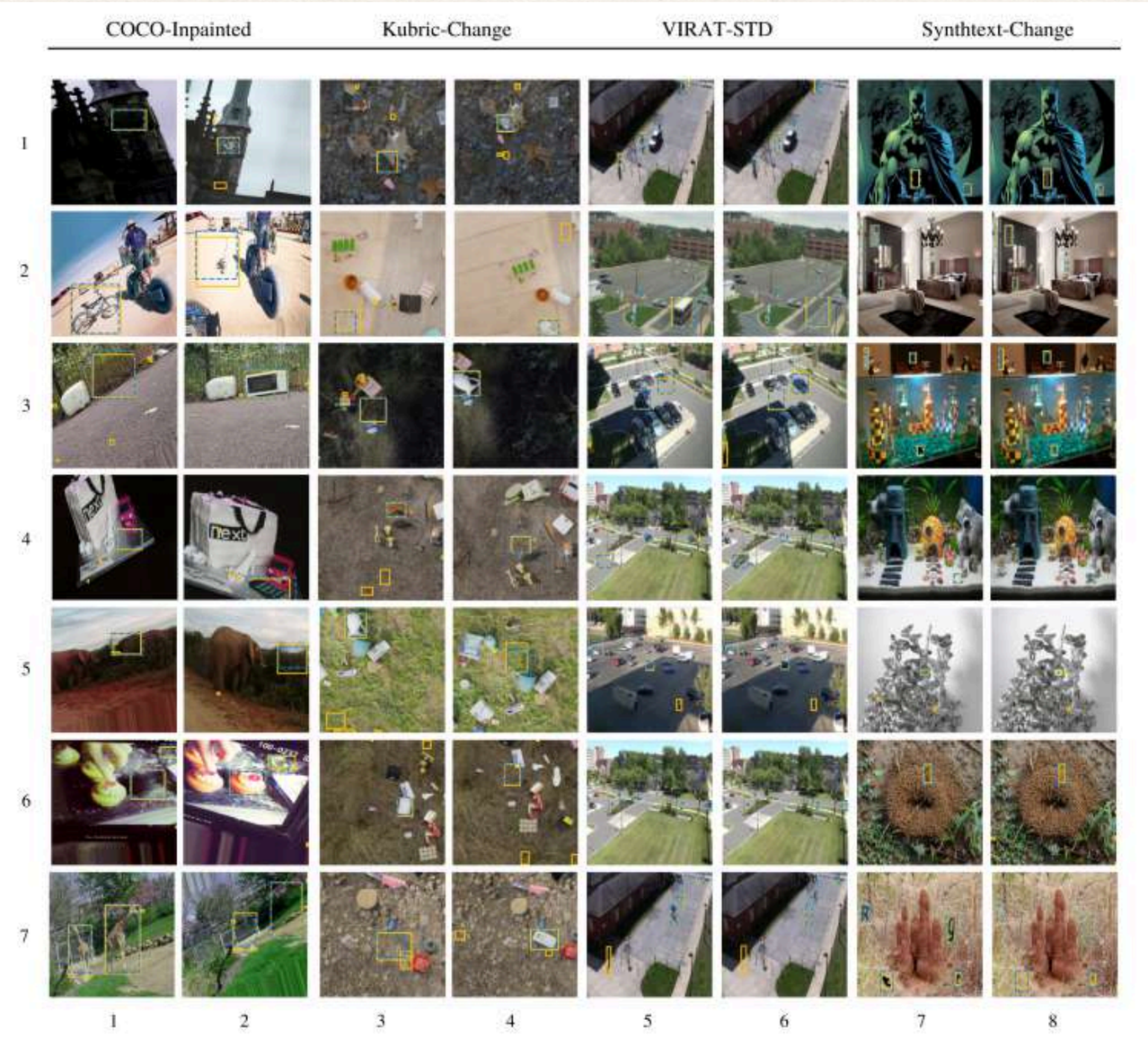


Figure 4. **Qualitative results:** We show the bounding box predictions (solid) of our model on all the test sets, along with the ground truth (dashed). Since the detection head outputs 100 bounding boxes per image (see Sec. 5.2), for the purpose of visualisation, we display the 5 most confident predictions. In case of multiple bounding boxes with significant overlap, we keep the most confident and suppress the others. Note the significant photometric changes in COCO-Inpainted, 3D geometric effects in Kubric-Change (notice the inside of the cup in row 2, col 3-4), detection of really small objects in VIRAT-STD (even picking up valid changes that are not part of the ground truth e.g. row 5, col 5-6) and very subtle letters in Synthtext-Change. We recommend that the reader zooms in on the individual image pairs for inspection.