FUSU: A Multi-temporal-source Land Use Change Segmentation Dataset for Fine-grained Urban Semantic Understanding

Shuai Yuan[1] **Guancong Lin**[2] **Lixian Zhang**[3] **Runmin Dong**[4] **Jinxiao Zhang**[4]
Shuang Chen[1] **Juepeng Zheng**[2*] **Jie Wang**[5] **Haohuan Fu**[6*]

[1]Department of Geography, The University of Hong Kong [2]School of Artificial Intelligence, Sun Yat-sen University [3] National Supercomputing Center in Shenzhen [4]Department of Earth System Science, Tsinghua University [5] Pengcheng Laboratory [6] Tsinghua Shenzhen International Graduate School, Tsinghua University
{shuai914, schen17}@connect.hku.hk; lingc8@mail2.sysu.edu.cn
zhanglx18@tsinghua.org.cn; drm@mail.tsinghua.edu.cn
zhang-jx22@mails.tsinghua.edu.cn; zhengjp8@mail.sysu.edu.cn
wangj10@pcl.ac.cn; haohuan@tsinghua.edu.cn

Abstract

Fine urban change segmentation using multi-temporal remote sensing images is essential for understanding human-environment interactions in urban areas. Although there have been advances in high-quality land cover datasets that reveal the physical features of urban landscapes, the lack of fine-grained land use datasets hinders a deeper understanding of how human activities are distributed across the landscape and the impact of these activities on the environment, thus constraining proper technique development. To address this, we introduce FUSU, the first fine-grained land use change segmentation dataset for Fine-grained Urban Semantic Understanding. FUSU features the most detailed land use classification system to date, with 17 classes and 30 billion pixels of annotations. It includes bi-temporal high-resolution satellite images with 0.2-0.5 m ground sample distance and monthly optical and radar satellite time series, covering 847 km^2 across five urban areas in the southern and northern of China with different geographical features. The fine-grained land use pixel-wise annotations and high spatial-temporal resolution data provide a robust foundation for developing proper deep learning models to provide contextual insights on human activities and urbanization. To fully leverage FUSU, we propose a unified time-series architecture for both change detection and segmentation. We benchmark FUSU on various methods for several tasks. Dataset and code are available at: *https://github.com/yuanshuai0914/FUSU*.

1 Introduction

Urban areas, housing 57% of the world's population on just 3% of global land, are dynamic hubs of human activity [1]. The scale and rapid pace of current urbanization, encompassing both internal dynamics and population growth, position urban areas as a crucial catalyst of global climate change and vice versa [2]. Therefore, proper observation and monitoring of urban changes are crucial for modeling human-nature interactions.

*Corresponding authors.

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

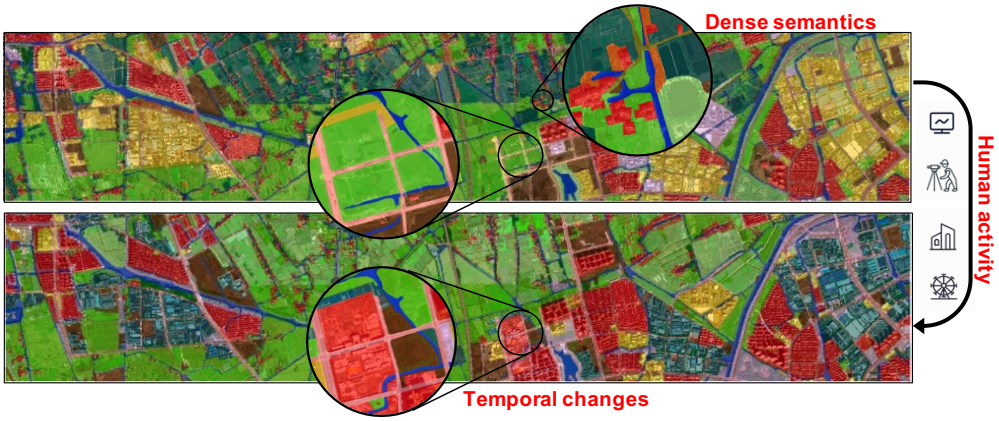

Figure 1: The unique features of urban areas. Compared with other geographic regions, urban areas have dense semantics, fast temporal changes, and involve a large amount of human activities.

In the era of data-driven methods, satellite remote sensing provides abundant data for Earth observation and deep learning-based models to comprehend the changes and mechanisms in such observations. However, urban areas have unique features requiring stringent conditions for high-quality data as Fig. 1 shows. First, multiple semantics are concentrated in small areas, and this dense semantic information is driven by human activities (land use) rather than natural characteristics (land cover) [3]. This necessitates high-resolution images and fine-grained land use annotations over land cover segmentation datasets. Second, urban areas undergo rapid temporal changes, demanding high-frequency observations to capture these dynamics accurately [4]. Third, Fig. 1 highlights the diversity of human activities during the urban changes, including work, construction, relocation, and entertainment, requiring multi-source data for effective monitoring.

Although numerous land cover change segmentation datasets (e.g., LoveDA [5], SECOND [6], Hi-UCD [7], DynamicEarthNet [8]) have been introduced to advance urban monitoring, their coarse-grained land cover classification systems still limit the ability of fine urban semantic understanding. For example, the SECOND dataset only focuses on six classes, including ground, trees, low vegetation, water, buildings, and playgrounds, which fails to capture the full range of urban elements and detailed land use information, thus inadequately reflecting urban conditions and human-urban interactions. Besides, due to the difficulties of acquiring multi-temporal high-resolution images (e.g., cloud obscuration, accessibility), most change segmentation datasets only comprise bi-temporal images with even single-temporal annotations, which cannot match the pace of urban development, leading to challenges in timely planning and management. A high spatial-temporal resolution change segmentation dataset with a fine land use classification system is required.

In this paper, we introduce FUSU, the first multi-temporal, multi-source land use change segmentation dataset with the finest pixel-wise change segmentation annotations to date, covering 17 land use classes and over 30 billion pixels. It includes bi-temporal high-resolution satellite images (0.2-0.5 m resolution) and aligns optical and radar satellite data (Sentinel-2, Sentinel-1) with monthly revisits, enriching temporal and multi-sensor information. Spanning 847 km^2 across five major urban districts in northern and southern China, FUSU's geographical diversity ensures domain shifts within the dataset. To leverage this spatial-spectral-temporal-resolution diversity, we propose FUSU-Net, a unified time-series architecture, as a baseline to make full utilization of the enriched information in FUSU for change detection and segmentation tasks. FUSU and FUSU-Net aim to advance dataset and algorithm development for improved urban monitoring and understanding. Our contributions include:

- We introduce FUSU, the first land use change segmentation dataset with a fine land use classification system of 17 classes and over 30 billion annotation pixels. FUSU captures timely urban dynamics from different perspectives and bridges the gaps between rich remote sensing data and urban semantic understanding.

- We showcase how the constructed time-series data can be leveraged for better urban monitoring by proposing a unified time-series baseline architecture FUSU-Net that conducts end-to-end change detection and segmentation tasks utilizing multi-temporal-source data.
- We benchmark FUSU on kinds of methods in several downstream tasks to provide a comprehensive insight.

2 Related Works

2.1 Urban Change Segmentation Data

Urban change segmentation is a critical aspect of Earth observation, garnering significant attention in recent years. Various land cover datasets have been developed to support specific tasks like change detection and segmentation (see Table 1). ISPRS Potsdam[2] provides high-resolution images for urban parsing, but it covers small areas and has a limited scale. SpaceNet [9], EuroSAT [10], and GID [11] cover larger areas but suffer from incomplete land cover classification, lower resolution, and single snapshots. LEVIR-CD [12] and WHU [13] focus on bi-temporal building change detection, but lack comprehensive semantics. SECOND [6], Hi-UCD [7], and WUSU [14] introduce multi-class semantic change detection. However, WUSU and Hi-UCD cover limited regions, and SECOND's coarse annotations and long intervals reduce continual observation capability. LoveDA [5] includes patches from various Chinese cities, but the classification system is coarse-grained, and the annotation only covers a single snapshot time. FLAIR [15] uses aerial and Sentinel-2 images for near-daily observations, yet only one temporal label cannot tell the changes during periods. DynamicEarthNet [8] provides daily observations and monthly annotations, but also suffers from the coarse-grained land cover classification system, which fails to provide semantics on human-environment interactions.

In summary, there is a lack of attention to land use datasets. Existing datasets usually present a trade-off among resolution, coverage, snapshot time, annotation pixel, and classification system. On the contrary, FUSU aims for the finest urban semantic understanding, providing the fine-grained land use classification system (17 classes), large-scale annotation pixels (30 billion), high-resolution images (0.2-0.5 m), large coverage (847 km^2), temporal information (bi-temporal high-resolution images and monthly Sentinel data), and supporting multiple downstream remote sensing tasks.

Table 1: A survey on open-source urban change segmentation datasets, including segmentation datasets and change detection datasets.

	Dataset	Source	Images (patches)	Size	Area (km^2)	Resolution (m)	Class	Objects	Temporal (image)	Temporal (annotation)	Ann pixel ($\times 10^9$)
Segmentation	Potsdam[1]	Aerial	38	6000	0.05	0.05	6	*LC*	1	1	0.8
	SpanceNet[9]	Maxar	60,000	650	5,500	0.3-1.24	2	*B&R*	1	1	1.3
	EuroSAT[10]	Sentinel-2	27,000	64	11,059	10	10	*LC*	1	1	0.1
	GID[11]	Gaofen-2	150	6800-7200	50,000	1	5/15	*LC*	1	1	7.3
	LoveDA[5]	Google Earth	5987	1024	536	0.3	6	*LC*	1	1	6.3
	FLAIR[15]	Aerial/Sentinel-2	77,762	512/40	817	0.2/10	18	*LC*	4 days	1	20.3
Change Detection	LEVIR-CD[12]	Google Earth	637	1024	167	0.5	1	*B*	2	1	0.005
	WHU[13]	Aerial	8,189	512	192	0.3	1	*B*	2	1	0.4
	SECOND[6]	Satellite	4,662	512	1,200	0.5-1	6	*LC*	2	2	0.9
	Hi-UCD[7]	Aerial	1,293	1024	30	0.1	9	*LC*	3	3	2.7
	WUSU[14]	Gaofen-2	2	5500-7025	80	1	11	*LC*	3	3	1.5
	DynamicEarthNet[8]	PlanetFusion	54,750	1024	16,986	3	7	*LC*	daily	monthly	1.9
	FUSU	**Google Earth/Sentinel-1/2**	**62,752**	**512/128**	**847**	**0.2-0.5/10**	**17**	***LU***	**monthly**	**2**	**32.2**

• *B*-Buildings, *R*-Roads, *LC*-Land Cover, *LU*-Land Use.

2.2 Remote sensing tasks

Change Detection identifies surface differences by processing images of the same area captured at different times [16]. It includes binary change detection [12, 13], which detects changes in a single class (changed or unchanged), and semantic change detection [6, 8, 17], which provides detailed land semantics. High-frequency observations are essential for timely geographical change detection, and fine-grained annotations improve precision. However, most datasets provide only bi-temporal

[2]https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx

observations due to the challenge of acquiring high-resolution multi-temporal images, resulting in long intervals that impede timely monitoring. Additionally, the lack of fine-grained multi-temporal annotations restricts the development of semantic change detection algorithms.

These challenges highlight the need for richer temporal data and fine-grained land use classifications, as well as methods capable of handling multi-temporal information. Current datasets' coarse-grained classifications do not accurately reflect urban conditions, and integrating multi-temporal data from other accessible sensors to enhance change detection has been underexplored. To address these issues, we propose FUSU, which includes bi-temporal fine-grained annotations and multi-temporal observations from high-resolution and Sentinel images. We also design a new unified architecture FUSU-Net to leverage time-series information for semantic change detection and segmentation.

Semantic segmentation has been widely applied in remote sensing for tasks such as land cover mapping [18], geographical object extraction [19, 20, 21], and cropland cover mapping [22]. Encoder-decoder architectures are well-suited to the diverse nature of remote sensing images [8]. Most studies focus on segmenting objects from static images [5, 23], while some have used time-series images to improve performance [22, 24]. Our FUSU-Net integrates time-series information into the bi-temporal segmentation task. We believe the unique time-series structure of FUSU will inspire the development of more advanced time-series segmentation algorithms in remote sensing.

3 FUSU Dataset

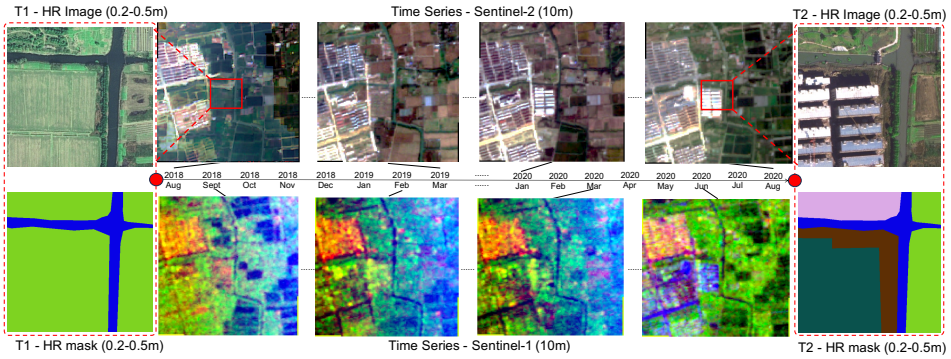

Figure 2: The visualization of the FUSU dataset construction. Each patch has 27 images (25 Sentinel images and 2 high-resolution images), and 2 labels. The content of the high-resolution image is center-surrounded by the Sentinel image as the red rectangle shows.

We introduce FUSU, a multi-temporal, multi-source change segmentation dataset for fine-grained urban semantic understanding. FUSU consists of 62,752 image patches, each containing 27 images from three sources with different resolutions and snapshot times, and includes two annotations as shown in Fig. 2. FUSU has four key features:

Fine-grained: FUSU features the finest land use classification system in change segmentation datasets, with bi-temporal dense annotations. It includes 17 classes—artificial-constructed, agricultural, and natural—that detail urban functional zoning and enhance understanding of urban structural development.

Multi-temporal: FUSU offers time-series observations with monthly revisits. Along with bi-temporal high-resolution images and fine-grained annotations, it supports high-frequency urban monitoring, enabling methods to leverage long-range temporal context for better inferences.

Multi-source: FUSU combines data from three satellite sources (Google Earth, Sentinel-2, Sentinel-1) with different temporal, resolution, and band compositions. Each image patch unifies spatial, temporal, and spectral contexts, providing richer information than single-source data.

Domain shifts: FUSU covers five urban areas in northern and southern China, each with diverse geographical features and urban landscapes. Variability in climate types and class ratios across these regions contribute to representation gaps and pronounced domain shifts in the feature data.

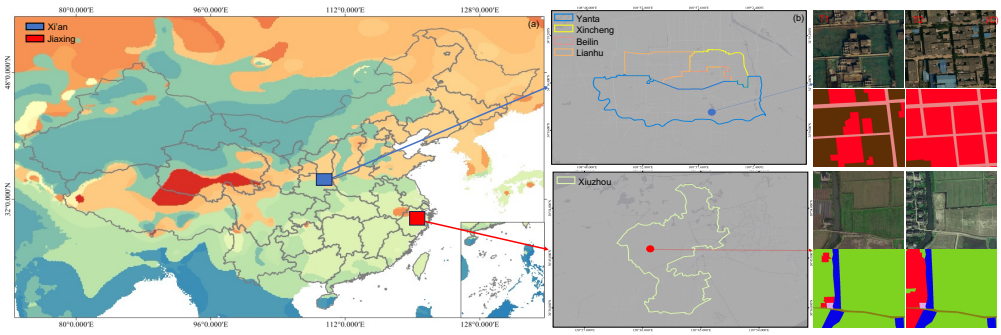

Figure 3: The distribution of the FUSU dataset. (a) Xi'an and Jiaxing are located in different climate zones. (b) The 5 urban districts of Xi'an and Jiaxing in FUSU dataset. (c) The visualization of image samples.

3.1 Construction of FUSU

Acquisitions. FUSU uses three data sources with different resolutions, geographical details, and acquisition times. Google Earth images are 512×512 pixels with a 0.3 m resolution and RGB bands. Sentinel-1 and Sentinel-2 images are sourced from Google Earth Engine (GEE). Sentinel-1 images are preprocessed by GEE (noise removal, radiometric calibration, orthorectification). Sentinel-2 images undergo cloud removal, atmospheric correction, radiometric calibration, and orthorectification, then are concatenated with Sentinel-1 data. Each Sentinel image is 128×128 pixels with a 10 m resolution and 14 bands. Google Earth and Sentinel patches are not strictly aligned; Google Earth patches cover only the central area of corresponding Sentinel patches (Fig. 2). This approach preserves semantic detail and captures broader context, aiding spatial dynamics understanding. More details are in the Sec. A.3.

Distribution. FUSU covers 847 km^2 across five urban districts in China: Xiuzhou in Jiaxing, and Yanta, Beilin, Xincheng, and Lianhu in Xi'an. The different climates of Jiaxing and Xi'an are illustrated in Fig. 3(a). FUSU provides continuous monthly observations from August 2018 to August 2020. Google Earth images were captured in August 2018 and August 2020, while Sentinel-1 and Sentinel-2 images were collected monthly between these dates.

Annotations. Bi-temporal Google Earth images are manually annotated pixel-wise by two teams of geography experts. Table 2 shows the classes, label values and colors. More details about the annotations can be found in Sec. A.1.

Table 2: Land use classification system of FUSU and corresponding label values, colors.

Color	Class	Label Value	Color	Class	Label Value	Color	Class	Label Value
	Traffic land	1		Industrial land	7		Special land	13
	Inland water	2		Orchard	8		Forest	14
	Residential land	3		Park	9		Storage	15
	Cropland	4		Public management	10		wetland	16
	Agriculture construction	5		Commercial land	11		Grass	17
	Blank	6		Public construction	12		Background	0

3.2 Statistic

FUSU includes bi-temporal pixel-level annotations covering 17 land use classes. Fig. 4(a) and (b) illustrate the distribution of pixels and polygons for each class at a single time snapshot. Residential land dominates both in terms of polygons and pixels. Some classes, like agriculture construction land, exhibit asymmetrical distributions. The highly unbalanced distribution numbers show a ratio exceeding 90 between the most and least frequent types. Fig. 4(c)-(f) display the class ratios in Xi'an and Jiaxing at two-time snapshots, revealing varying distributions between the cities. Jiaxing is characterized by significant cropland and residential areas, while Xi'an has more commercial land. These class imbalances and city differences pose challenges for urban monitoring using FUSU.

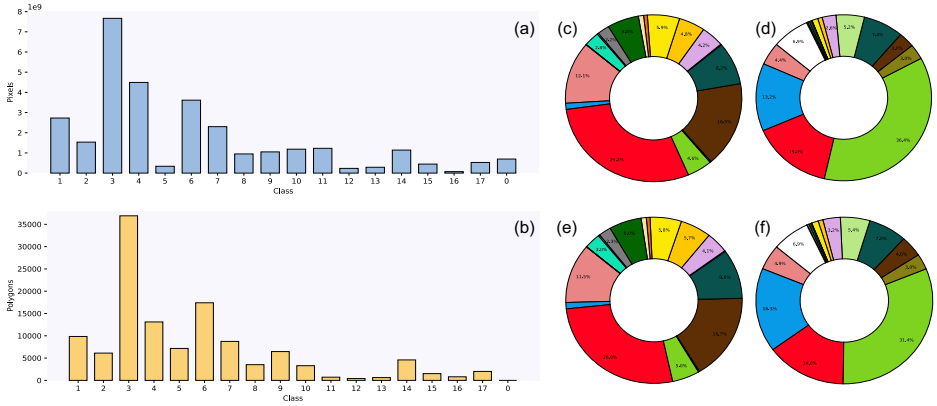

Figure 4: The statistic of the FUSU dataset. (a) Pixels distribution. (b) Polygon distribution. (c) Class distribution of T1 Xi'an. (d) Class distribution of T1 Jiaxing. (e) Class distribution of T2 Xi'an. (f) Class distribution of T2 Jiaxing.

4 FUSU-Net

To fully utilize FUSU, we propose a unified time-series baseline architecture named FUSU-Net that conducts end-to-end change detection and segmentation tasks. Fig. 5 shows the architecture.

4.1 Preliminary and Overview

Given T1 image $\mathcal{I}_1$, T2 image $\mathcal{I}_2$, the corresponding groundtruth labels $\mathcal{Y}_1$, $\mathcal{Y}_2$, and the time-series temporal images $\mathcal{I}_T$, we have two ultimate goals: build a segmentation function $\mathcal{F}_s$ that generates segmentation map $\hat{\mathcal{Y}} = \mathcal{F}_s(\mathcal{I} \mid \mathcal{I}_T)$, and build a change detection function $\mathcal{F}_c$ that find binary changes between two input images $\hat{\mathcal{C}} = \mathcal{F}_c(\hat{\mathcal{Y}}_1, \hat{\mathcal{Y}}_2 \mid \mathcal{I}_T)$. These two goals mean we need to optimize the loss $\mathcal{L}$ between predicted values and labels:

$$\theta^* = \arg\min_{\theta}\{\mathcal{L}^s(\mathcal{F}_s(\mathcal{I} \mid \mathcal{I}_T), \mathcal{Y}) + \mathcal{L}^c(\mathcal{F}_c(\hat{\mathcal{Y}}_1, \hat{\mathcal{Y}}_2 \mid \mathcal{I}_T), \mathcal{Y}_c)\}, \tag{1}$$

where θ^* is the optimized learned parameters generated by the optimized $\mathcal{L}^c$ and $\mathcal{L}^s$, and θ represents the learned parameters, and $\mathcal{Y}_c$ is the binary change groundtruth label, which can be generated by $\mathcal{Y}_1$, $\mathcal{Y}_2$:

$$y_c^{(i,j)} = \begin{cases} 0, & y_1^{(i,j)} = y_2^{(i,j)} \\ 1, & y_1^{(i,j)} \neq y_2^{(i,j)}, \end{cases} \tag{2}$$

where $y^{(i,j)}$ is the pixel value. Assuming the additional temporal and spectral information in time-series images can guide the high-resolution segmentation and change detection, we further extract the high-level temporal and spectral information and use $\mathcal{Y}_1$ for supervision. Thus the optimization body can be divided into:

$$\begin{aligned} \theta^* = \arg\min_{\theta}\{&\mathcal{L}_1^s(\mathcal{F}_s(\mathcal{I} \mid \mathcal{F}_s(\mathcal{I}_T, \mathcal{Y}_1; \theta), \mathcal{Y}_1; \theta)) + \mathcal{L}_2^s(\mathcal{F}_s(\mathcal{I} \mid \mathcal{F}_s(\mathcal{I}_T, \mathcal{Y}_1; \theta), \mathcal{Y}_2; \theta)) + \\ &\mathcal{L}_T^s(\mathcal{F}_s(\mathcal{I}_T, \mathcal{Y}_1; \theta)) + \mathcal{L}^c(\mathcal{F}_c(\hat{\mathcal{Y}}_1, \hat{\mathcal{Y}}_2 \mid \mathcal{F}_s(\mathcal{I}_T, \mathcal{Y}_1; \theta)), \mathcal{Y}_c; \theta)\}, \end{aligned} \tag{3}$$

where $\mathcal{L}^s_{\{1,2,T\}}$ is the loss of segmentation of T1 image, T2 image, and time-series images, respectively.

4.2 Overall architecture

As Fig. 5 shows, the overall architecture of FUSU-Net includes two branches: (a) This branch processes Sentinel time-series images and outputs time-series features; (b) This branch processes

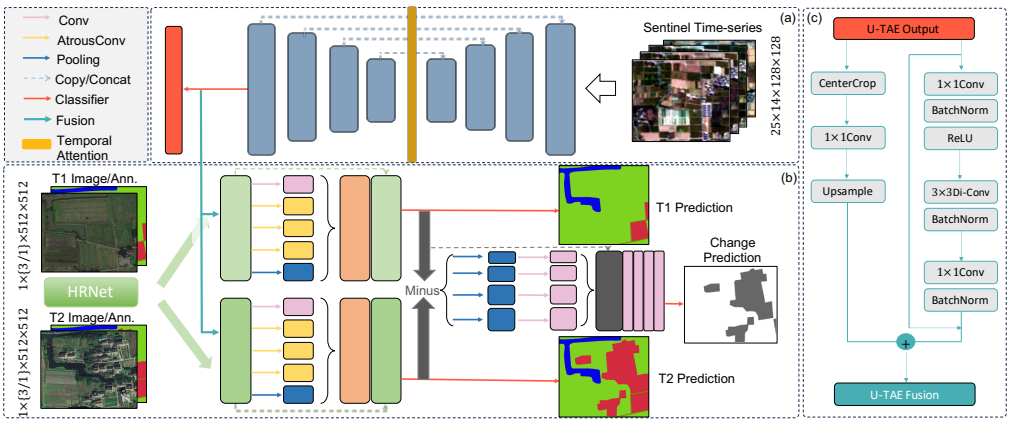

Figure 5: The architecture of FUSU-Net. (a) U-TAE branch for time-series images. (b) Bi-temporal branch for segmentation and change detection. (c) Feature fusion.

bi-temporal high-resolution images and annotations and outputs both bi-temporal segmentation results and change detection results.

As Fig. 5(a) shows, to process the Sentinel time-series images, we use U-TAE [22] with temporal attention to effectively capture temporal information in feature maps at various resolutions. The input shape is $25 \times 14 \times 512 \times 512$ ($T \times C \times H \times W$) and the output shape is $64 \times 512 \times 512$. Fig. 5(b) shows that we first use an HR-Net pre-trained on ImageNet as the backbone to extract bi-temporal features. Then we input each feature into separated ASPP [25] segmentation heads to get the segmentation results. We then conduct a minus operation between bi-temporal segmentation features, and after a Spatial Pyramid Pooling head [26], we can get the binary change detection result. Note that Fig. 5(c) shows the fusion module. Time-series features fuse with bi-temporal features via two transformations: First, the time-series feature is center-cropped to strictly geographically align with the bi-temporal features. Then after a 1×1 convolution and upsampling layer, the center-cropped feature has the same shape with bi-temporal features. Second, we reserve the large spatial information of the time-series feature and after a bottle-neck structure with a dilated convolution, we map it to the same shape of the bi-temporal features. An add operation is conducted for the feature fusion.

4.3 Loss Functions

As discussed in Sec. 4.1, we use 4 loss functions to train FUSU-Net: three segmentation loss $\mathcal{L}^s_{\{1,2,T\}}$, and a change loss $\mathcal{L}^c$. The segmentation loss functions are the multi-class cross-entropy loss. Specifically, for time-series supervision, we first centercrop the output for geographical alignment, then upsample it to the same size of groundtruth label $\mathcal{Y}_1$. The change loss is the BCE loss to supervise the binary changes. More details about supervision and implementation can be found in Sec. A.5.3.

5 Experiments

We utilize our dataset for semantic segmentation in Sec. 5.1 and change detection in Sec. 5.2 with various experiments on state-of-the-art baseline methods and FUSU-Net. We also validate the feature disparities between Jiaxing and Xi'an in the segmentation task.

5.1 Semantic Segmentation

Land use segmentation is crucial for urban monitoring. We focus on single-temporal images and labels for this semantic segmentation task. We compare other seven baseline segmentation methods with our FUSU-Net: FCN [27], PSPNet [26], Fast-SCNN [28], Deeplab-v3 [25], HRNet [29], K-net [30], and U-TAE [22]. Evaluation is based on intersection over union (IoU) per class and mean IoU (mIoU) across all 17 land use classes, following established protocols. Additionally, we investigate feature disparities between Jiaxing and Xi'an through two experiments: intra-dataset (whole, Xi'an,

Jiaxing) and inter-dataset (training on one, testing on the other). Implementation details are provided in the Sec. A.5.2.

Table 3: Semantic segmentation results obtained from intra-dataset.

Method	IoU per class (%)																	mIoU
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	
FCN [27]	70.83	76.49	74.67	84.14	30.84	**52.16**	53.39	33.32	**52.7**	**50.24**	28.98	0.09	30.62	57.42	23.61	13.04	17.79	44.25
PSPNet [26]	65.37	79.15	71.44	82.354	23.5	49.97	52.82	40.44	44.90	44.50	31.39	30.08	24.04	48.69	41.64	24.50	32.58	46.32
Fast-SCNN [28]	54.42	72.28	66.25	78.97	2.80	42.40	47.84	35.33	30.24	30.24	31.94	12.03	0	0	**44.66**	31.75	23.39	35.56
Deeplab-v3 [25]	66.17	77.31	71.20	82.10	26.30	49.61	53.96	37.35	45.85	47.61	33.21	**35.06**	**30.68**	54.14	34.94	**34.15**	32.07	47.74
HRNet [29]	67.6	**80.39**	73.24	83.02	22.94	49.00	54.05	40.10	46.43	49.12	31.66	26.68	15.21	52.38	42.84	30.16	32.09	46.88
K-net [30]	59.97	72.68	66.87	79.46	18.45	44.19	48.07	32.05	35.07	35.64	19.9	18.44	18.69	49.61	29.2	23.88	22.51	39.69
U-TAE [22]	59.57	64.18	65.76	77.92	24.87	40.13	46.75	29.89	41.72	30.57	26.13	6.85	25.96	30.57	41.83	15.08	8.12	37.63
FUSU-Net	**74.79**	78.95	**76.13**	**85.35**	**34.81**	50.54	**53.47**	**41.50**	49.64	45.78	**36.69**	28.85	28.98	**60.21**	44.41	30.07	**33.69**	**50.10**

Table 4: Semantic segmentation results obtained from inter-dataset.

Method	mIoU			
	Training on Xi'an Testing on Xi'an	Training on Jiaxing Testing on Jiaxing	Training on Xi'an Testing on Jiaxing	Training on jiaxing Testing on Xi'an
FCN [27]	50.21	45.53	9.07	9.36
PSPNet [26]	46.52	43.35	8.55	9.72
Fast SCNN [28]	32.97	32.76	7.83	8.51
HRNet [29]	46.78	45.01	10.07	9.73
K-net [30]	38.17	37.41	9.31	**10.59**
Deeplab-v3 [25]	47.30	47.89	9.65	9.13
FUSU-Net	**53.63**	**49.91**	**11.65**	10.46

Overall results. Table 3 shows the segmentation results. We observe that FUSU-Net achieves the best results regarding mIoU. Specifically, FUSU-Net performs better than other methods not only on some comparatively simple classes (i.e., traffic land-1, residential land-3) but also has continuous promising results on difficult classes where other methods have poor performance (i.e., commercial land-11, special land-13). Note that FUSU-Net is backboned by HRNet and the segmentation head is PSPNet with FCN, and the results directly show the benefits of adding features of time-series Sentinel images. When compared with U-TAE, we can see that high-resolution images can also improve performance by providing more clear observation details.

Cross-dataset results. Table 4 shows the segmentation results with different training and testing datasets. There is a dramatic drop in mIoU on cross-dataset training and testing compared with training and testing on the same datasets. We can tell the huge feature differences between Jiaxing and Xi'an from these results.

Table 5: Semantic change detection results obtained from intra-dataset.

Method	IoU per class (%)																	mIoU
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	
BIT [31]	35.54	48.90	46.89	42.27	4.01	46.70	59.92	23.6	35.41	25.82	17.88	0	3.95	54.23	22.62	12.20	46.15	30.95
ChangeFormer [32]	39.31	57.87	57.13	39.42	9.20	25.58	60.11	31.33	27.17	19.79	12.07	0.31	7.42	59.81	19.71	**35.61**	35.13	32.17
ICIFNet [33]	49.75	56.41	62.23	51.21	4.7	53.81	**61.43**	30.03	**47.35**	3.47	10.45	0	0	**73.65**	53.18	11.15	60.75	36.17
DMINet [34]	26.63	34.08	54.91	42.75	0	32.59	39.80	17.91	19.34	0	6.35	0	0	21.83	39.41	20.94	54.18	24.16
SSCD-l [35]	23.19	15.95	31.32	29.12	6.12	35.46	27.08	12.30	18.91	2.50	0	0	3.39	2.06	20.51	16.34	15.69	15.29
Bi-SRNet [35]	26.19	41.42	39.82	40.01	**21.18**	44.26	46.59	26.70	25.05	**31.23**	**20.21**	**7.18**	4.74	40.91	31.66	30.87	37.40	30.91
FUSU-Net	**55.67**	**61.46**	**66.19**	**55.83**	19.82	**55.22**	57.86	**34.59**	46.43	15.45	16.31	5.89	**9.47**	65.12	**54.32**	14.45	**64.52**	**41.09**

5.2 Change Detection

We then compare the performance of change detection baselines on FUSU. Here, we complete the binary change detection (BCD) experiment and semantic change detection (SCD) experiment. For binary change detection, we introduce 6 methods: DMINet [34], ICIFNet [33], ChangeFormer [32], A2Net [36], BIT [31], USSFC-Net [37]. We evaluate the results by IoU on changed pixels. For semantic change detection, we introduce 6 methods: BIT [31], ChangeFormer [32], ICIFNet [33], DMINet [34], SSCD-l [35], Bi-SRNet [35]. We evaluate the change detection results by IoU per class and mIoU over all 17 land use classes. Implementation details can be found in Sec. A.5.2.

Overall results. Table 6 and Table 5 present the results of binary and semantic change detection. In binary change detection, with only unchanged and changed pixels, class-specific IoU is not applicable. Our FUSU-Net outperforms other baselines by 7.21%-31.89% in IoU. In semantic change detection, challenging classes such as public management-10, public construction-12, and special land-13 are observed across all methods, consistent with semantic segmentation results. Notably, FUSU-Net achieves better performance compared to other baseline methods than it does in the semantic segmentation task, which can be attributed to continuous observation and change information provided by time-series Sentinel images between two high-resolution image snapshots.

Table 6: Binary change detection results obtained from intra-dataset.

Method	IoU	Method	IoU
BIT [31]	47.91	ChangeFormer [32]	59.64
ICIFNet [33]	64.74	DMINet [34]	72.59
A2Net [36]	69.22	USSFC-Net [37]	62.85
FUSU-Net	79.80		

6 Discussion

Table 7: Ablation results on the effectiveness of time-series.

Time-series	0	9	18	25
mIoU (Seg)	46.72	47.19	48.47	50.10
IoU (BCD)	65.51	69.35	74.39	79.80
mIoU (SCD)	26.64	34.14	36.55	41.09

Effectiveness of time-series. We evaluate to what extent time-series images enhance the performance. Table 7 shows the results. We choose the number of time-series images as the variable (i.e., all time-series images, partial time-series images, zero time-series images). We can see for the FUSU dataset, more time-series images contribute to better results. It is desirable to use all time-series images as additional temporal information.

Limitations. The FUSU dataset has three primary limitations. First, it is limited to five urban districts. Despite its rich geographical diversity and pixel data, including more global urban areas is desirable. We encourage the community to share high-quality, fine-grained land use datasets to advance urban monitoring. Second, land use change segmentation requires understanding human activities and production, unlike land cover, which directly corresponds to pixel values. Relying solely on remote sensing imagery makes high accuracy challenging. Third, as Table 8 shows, because of the sensor gaps between optical images and SAR images, the simple concatenation of Sentinel-1 and Sentinel-2 is not ideal. Better fusion methods should be considered for synergizing both Sentinel-2 and Sentinel-1 strengths. In the future, we aim to design optical-SAR fusion methods and incorporate more multi-source data, such as economic and population data, to develop a multi-modal framework for comprehensive urban semantic understanding.

Table 8: Effectiveness of Sentinel-1.

	+ S1	- S1
mIoU (Seg)	50.10	51.17

Conclusion. We present FUSU, a comprehensive multi-source, multi-temporal change segmentation dataset for fine-grained urban semantic understanding. FUSU includes a detailed 17-class land use classification system, 30 billion annotated pixels, 847 km² coverage, and temporal information from bi-temporal high-resolution images and monthly Sentinel data. This makes FUSU the most comprehensive urban semantic dataset available. We benchmark various methods to demonstrate FUSU's effectiveness in urban land use segmentation and change detection. Additionally, we introduce FUSU-Net, a model that fully utilizes the spatial, spectral, and temporal diversity of FUSU. We anticipate that FUSU and FUSU-Net will advance the development of powerful techniques for multi-source, multi-temporal change segmentation in urban environments without any negative societal impacts.

7 Acknowledgement

We acknowledge the annotation team for their high-quality work. This work was supported in part by the National Key Research and Development Plan of China (Grant No. 2023YFB3002400), and in part by the National Natural Science Foundation of China (Grant T2125006 and No. 42401415), and in part by Jiangsu Innovation Capacity Building Program (Project No. BM2022028).

References

[1] Karen C Seto, Burak Güneralp, and Lucy R Hutyra. Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools. *Proceedings of the National Academy of Sciences*, 109(40):16083–16088, 2012.

[2] Karen C Seto, Anette Reenberg, Christopher G Boone, Michail Fragkias, Dagmar Haase, Tobias Langanke, Peter Marcotullio, Darla K Munroe, Branislav Olah, and David Simon. Urban land teleconnections and sustainability. *Proceedings of the National Academy of Sciences*, 109(20): 7687–7692, 2012.

[3] Fabio Pacifici, Marco Chini, and William J Emery. A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban land-use classification. *Remote Sensing of Environment*, 113(6):1276–1292, 2009.

[4] Amarasinghaghe Tharindu Dasun Perera, Kavan Javanroodi, Dasaraden Mauree, Vahid M Nik, Pietro Florio, Tianzhen Hong, and Deliang Chen. Challenges resulting from urban density and climate change for the eu energy transition. *Nature Energy*, pages 1–16, 2023.

[5] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021.

[6] 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.

[7] Shiqi Tian, Ailong Ma, Zhuo Zheng, and Yanfei Zhong. Hi-ucd: A large-scale dataset for urban semantic change detection in remote sensing imagery. *arXiv preprint arXiv:2011.03247*, 2020.

[8] Aysim Toker, Lukas Kondmann, Mark Weber, Marvin Eisenberger, Andrés Camero, Jingliang Hu, Ariadna Pregel Hoderlein, Çağlar Şenaras, Timothy Davis, Daniel Cremers, et al. Dynamicearthnet: Daily multi-spectral satellite dataset for semantic change segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21158–21167, 2022.

[9] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018.

[10] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

[11] Xin-Yi Tong, Gui-Song Xia, Qikai Lu, Huanfeng Shen, Shengyang Li, Shucheng You, and Liangpei Zhang. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sensing of Environment*, 237:111322, 2020.

[12] 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.

[13] Shunping Ji, Shiqing Wei, and Meng Lu. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on geoscience and remote sensing*, 57(1):574–586, 2018.

[14] Sunan Shi, Yanfei Zhong, Yinhe Liu, Jue Wang, Yuting Wan, Ji Zhao, Pengyuan Lv, Liangpei Zhang, and Deren Li. Multi-temporal urban semantic understanding based on gf-2 remote sensing imagery: from tri-temporal datasets to multi-task mapping. *International Journal of Digital Earth*, 16(1):3321–3347, 2023.

[15] Anatol Garioud, Nicolas Gonthier, Loic Landrieu, Apolline De Wit, Marion Valette, Marc Poupée, Sébastien Giordano, et al. Flair: a country-scale land cover semantic segmentation dataset from multi-source optical imagery. *Advances in Neural Information Processing Systems*, 36, 2024.

[16] Wenzhong Shi, Min Zhang, Rui Zhang, Shanxiong Chen, and Zhao Zhan. Change detection based on artificial intelligence: State-of-the-art and challenges. *Remote Sensing*, 12(10):1688, 2020.

[17] Shuai Yuan, Fan Wei, Lixian Zhang, Haohuan Fu, and Peng Gong. Receptive convolution boosts large-scale multi-class change detection. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 10459–10462. IEEE, 2024.

[18] Runmin Dong, Lichao Mou, Mengxuan Chen, Weijia Li, Xin-Yi Tong, Shuai Yuan, Lixian Zhang, Juepeng Zheng, Xiaoxiang Zhu, and Haohuan Fu. Large-scale land cover mapping with fine-grained classes via class-aware semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16783–16793, 2023.

[19] Shuai Yuan, Runmin Dong, Juepeng Zheng, Wenzhao Wu, Lixian Zhang, Weijia Li, and Haohuan Fu. Long time-series analysis of urban development based on effective building extraction. In *Geospatial Informatics X*, volume 11398, pages 192–199. SPIE, 2020.

[20] Lixian Zhang, Shuai Yuan, Runmin Dong, Juepeng Zheng, Bin Gan, Dengmao Fang, Yang Liu, and Haohuan Fu. Swcare: Switchable learning and connectivity-aware refinement method for multi-city and diverse-scenario road mapping using remote sensing images. *International Journal of Applied Earth Observation and Geoinformation*, 127:103665, 2024.

[21] Shuai Yuan, Juepeng Zheng, Lixian Zhang, Runmin Dong, Yile Xing, Yuhan She, Haohuan Fu, and Ray CC Cheung. Melting glacier: A 37-year (1984–2020) high-resolution glacier-cover record of mt. kilimanjaro. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 3943–3946. IEEE, 2022.

[22] Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4872–4881, 2021.

[23] Caleb Robinson, Le Hou, Kolya Malkin, Rachel Soobitsky, Jacob Czawlytko, Bistra Dilkina, and Nebojsa Jojic. Large scale high-resolution land cover mapping with multi-resolution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12726–12735, 2019.

[24] Michail Tarasiou, Erik Chavez, and Stefanos Zafeiriou. Vits for sits: Vision transformers for satellite image time series. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10418–10428, 2023.

[25] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[26] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[27] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[28] Rudra PK Poudel, Stephan Liwicki, and Roberto Cipolla. Fast-scnn: Fast semantic segmentation network. *arXiv preprint arXiv:1902.04502*, 2019.

[29] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10): 3349–3364, 2020.

[30] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. *Advances in Neural Information Processing Systems*, 34:10326–10338, 2021.

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

[32] Wele Gedara Chaminda Bandara and Vishal M Patel. A transformer-based siamese network for change detection. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 207–210. IEEE, 2022.

[33] Yuchao Feng, Honghui Xu, Jiawei Jiang, Hao Liu, and Jianwei Zheng. Icif-net: Intra-scale cross-interaction and inter-scale feature fusion network for bitemporal remote sensing images change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022.

[34] Yuchao Feng, Jiawei Jiang, Honghui Xu, and Jianwei Zheng. Change detection on remote sensing images using dual-branch multilevel intertemporal network. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.

[35] Lei Ding, Haitao Guo, Sicong Liu, Lichao Mou, Jing Zhang, and Lorenzo Bruzzone. Bi-temporal semantic reasoning for the semantic change detection in hr remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022.

[36] Zhenglai Li, Chang Tang, Xinwang Liu, Wei Zhang, Jie Dou, Lizhe Wang, and Albert Y Zomaya. Lightweight remote sensing change detection with progressive feature aggregation and supervised attention. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–12, 2023.

[37] Tao Lei, Xinzhe Geng, Hailong Ning, Zhiyong Lv, Maoguo Gong, Yaochu Jin, and Asoke K Nandi. Ultralightweight spatial–spectral feature cooperation network for change detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023.

[38] Haoran Wang, Tong Shen, Wei Zhang, Ling-Yu Duan, and Tao Mei. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In *European conference on computer vision*, pages 642–659. Springer, 2020.

[39] Qing Lian, Fengmao Lv, Lixin Duan, and Boqing Gong. Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6758–6767, 2019.

[40] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2507–2516, 2019.

[41] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018.

[42] 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.

[43] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 415–430. Springer, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Sec. 6 Limitations part.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Sec. 6 conclusion part and Supplementary Materials. We do not foresee any negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We have put the data, code, and instructions into the Github link mentioned in the Abstract.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Due to the page limits, we put them into the Appendix/Supplementary Materials, Sec. A.5.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Refer to Appendix Sec. A.5.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider? [Yes] These are included along with the training details in Appendix/Supplementary Materials, Sec. A.5
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Sec. 3 about Sentinel data, Sec. 4 about U-TAE.
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See the Abstract.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Sec. 3.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Sec. 6
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix

A.1 Annotations

The 17 land use classes used in FUSU are annotated according to the Chinese Land Use Classification Criteria (GB/T21010-2017) Level-1 classification system, i.e., *traffic land, inland water, residential land, cropland, agriculture construction, blank, industrial land, orchard, park, public management, commercial land, public construction, special land, forest, storage, wetland, grass, background*. The detailed criteria and description of each class are shown in Table 9. The annotation is conducted by two teams of geo-experts based on the ArcGIS geospatial software. Each team is responsible for one city and the annotation results are cross-checked by the other team. If there exists disagreement in some areas, these areas will be re-annotated when the agreement is reached. Leaders of two teams will randomly select 100 small areas in two cities for quality check. All objects are annotated as polygon features. Total annotation costs about 3 months. To ensure geographical continuity, the annotation is conducted on the full-scale images before image cropping.

Table 9: Class description and criteria.

Value	Name	Criteria
1	***Traffic land***	Refers to land for transportation facilities and their affiliated facilities such as railways, highways, airports, ports, docks, pipelines, urban rail transit, various roads, and transport stations, excluding auxiliary roads and parking lots within other lands
2	***Inland water***	Refers to natural land water bodies within the land area such as rivers, lakes, glaciers, and perennial snow, as well as artificial land water bodies such as reservoirs, ponds, and canal water surfaces
3	***Residential land***	Refers to urban and rural residential land and land for community service facilities supporting residential life
4	***Cropland***	Refers to land mainly used for cultivating crops, with at least one crop cycle per year (including land used for perennial crops cultivated in a manner of one or more crop cycles per year). This includes mature land, newly developed, reclaimed, and organized land, fallow land (including fallow rotation and fallow land)
5	***Agriculture construction***	Refers to land where the surface cultivation layer has been destroyed for the service of agricultural production and rural life, including rural roads and construction land for planting facilities, livestock and poultry facilities, and aquaculture facilities
6	***Blank***	Refers to land within urban and village areas designated by national space planning with unclear planning use, not to be developed within the planning period or to be developed under specific conditions
7	***Industrial land***	Refers to land used for industrial and mining production
8	***Orchard***	Refers to land used for cultivating perennial crops intensively for the collection of fruits, leaves, roots, stems, or sap, with a coverage rate of more than 50% or more than 70% of the reasonable number of plants per acre, including land used for nurseries
9	***Park***	Refers to land within urban and village construction areas for parks, protective green spaces, squares, and other public open spaces, excluding auxiliary green spaces in other construction lands
10	***Public management***	Refers to land for institutions and facilities of administrative bodies, groups, research, culture, education, sports, health, social welfare, etc., excluding rural and urban community service facilities
11	***Commercial land***	Refers to land for commercial, business finance, and recreational facilities, excluding rural and urban community service facilities
12	***Public construction***	Refers to land for urban and regional infrastructure facilities such as water supply, drainage, power supply, gas supply, heating, communication, postal services, broadcasting, sanitation, firefighting, main channels, and hydraulic works
13	***Special land***	Refers to land for military, foreign affairs, religious, security, funeral purposes, and sites of historical relics with special properties
14	***Forest***	Refers to land growing trees, bamboo, or shrubs. This does not include wetland growing trees, greening land within urban and village areas, trees within the scope of railway and highway land, or trees for river and canal embankment protection
15	***Storage***	Refers to land for logistics storage and strategic material reserve warehouses
16	***Wetland***	Refers to the land at the interface of land and water bodies where the water level is close to or at the surface, or with shallow water layers, remaining in a natural state
17	***Grass***	Refers to land mainly growing herbaceous plants, including sparse forest grasslands with a tree canopy density of less than 0.1 and shrub grasslands with shrub coverage of less than 40%. This does not include wetlands or saline-alkali lands growing herbaceous plants
0	***Background***	Others or extremely difficult to annotate

A.2 Sentinel Time Series

Sentinel-2. The Sentinel-2 sensor is a multispectral sensor launched in 2015. The Sentinel-2 we use has 12 bands covering the VNIR and SWIR regions, with spatial resolutions of 10, 20, and 60 m. The swath width is 290 km. In general, the complete survey of the earth is repeated every 5 days. Here, we select all available Level-2A products (Bottom-Of-the-Atmosphere reflectances) in one single month, which are preprocessed through atmosphere correction, and compute the mean of these products to get the monthly-revisited observation data. All images are cloud-free by s2cloudless[3]. Table 10 summarizes the spectral and spatial attributes and applications of Sentinel-2 bands. Note that Sentinel-2 sensors have 10, 20, and 60 m spatial resolutions, and all bands are resampled to 10 m by the nearest interpolation method.

[3]https://github.com/sentinel-hub/sentinel2-cloud-detector

Table 10: Spectral and spatial attributes of Sentinel-2.

Original band number	FUSU band number	Band width (mm)	Center band (mm)	Original resolution (m)	FUSU resolution (m)	Usage
1	1	20	443	60	10	Atmospheric correction
2	2	65	490	10	10	Vegetation aerosol scattering
3	3	35	560	10	10	Green peak
4	4	30	665	10	10	Max chlorophyll absorption
5	5	15	705	20	10	Not used in L2A context
6	6	15	740	20	10	Not used in L2A context
7	7	20	783	20	10	Not used in L2A context
8	8	115	842	10	10	LAI
8a	9	20	865	20	10	Water vapor absorption reference
9	10	20	945	60	10	Water vapor absorption atmospheric correction
11	11	90	1610	20	10	Soils detection
12	12	180	2190	20	10	AOT determination

Sentinel-1. The Sentinel-1 mission provides data from a dual-polarization C-band Synthetic Aperture Radar (SAR) instrument at 5.405GHz (C band). The satellites are to operate day-and-night and perform a synthetic aperture with radar imaging in all weather conditions. Sentinel-1 images in FUSU have 2 bands VV and VH (dual-band cross-polarization, vertical transmit/horizontal receive). The revisited cycle is 6 days. To get the monthly-revisited data, we also first select all available products in one single month and process the raw data by noise removal, radiometric calibration, and orthorectification, and then compute the mean of these products.

A.3 Dataset Expanding and Benefits

Dataset expanding. To expand the temporal information of FUSU, we develop a data-expanding paradigm that combines the temporally rich Sentinel images with high-resolution Google Earth images. This method involves three steps. First, we crop Google Earth images into 512×512 patches and generate a shapefile that is five times larger centered around the patch's midpoint for each patch. Then these shapefiles are used to download Sentinel images from Google Earth Engine. The time series of Sentinel images span the entire time interval between the snapshot times of two Google Earth images, with one image per month. Sentinel-1 images are preprocessed by noise removal, calibration, and correction. We then process Sentinel-2 to usable conditions by cloud removal, atmosphere correction, radiometric calibration, and orthorectification. As a result, Sentinel images have a size of 128×128 with a resolution of 10 m. Note that the Sentinel patches and Google Earth patches are not strictly aligned. The geographic content covered by a Google Earth patch only occupies the central areas of the corresponding Sentinel patch as shown in Fig. 6. This consideration is adopted for two main reasons: First, if strict alignment were enforced, the Sentinel patch size would be very small due to the significant resolution difference, resulting in insufficient semantic information for model training. Second, the larger coverage area of Sentinel patches captures the surrounding context and landscape variations and helps identify and understand patterns and trends in broader spatial dynamics.

This data-expanding paradigm enhances FUSU's temporal resolution, capturing more detailed changes during time series. Moreover, it is versatile enough to be extended to other readily available change detection datasets. We will provide the process steps and code accordingly for the community.

Benefits. Supplementing bi-temporal high-resolution images with the public multi-temporal Sentinel-2 and Sentinel-1 images has benefits in both clear geographic feature awareness and feature change awareness. First, Sentinel images provide high temporal-resolution observations, filling the gap of continuous temporal information between the snapshot times of bi-temporal images. This enables the capture of monthly changes and enhances the ability to detect and understand changes over time. Second, Sentinel images offer extensive spatial information. Thanks to the data-expansion design described in Sec. A.3, each Sentinel image is centered around the corresponding high-resolution image. This additional spatial information provides larger receptive fields for our model, ensuring geographical continuity and enabling broader area observations. Third, Sentinel-1 and Sentinel-2 images provide diverse observations from different modalities. These varied modalities enhance the

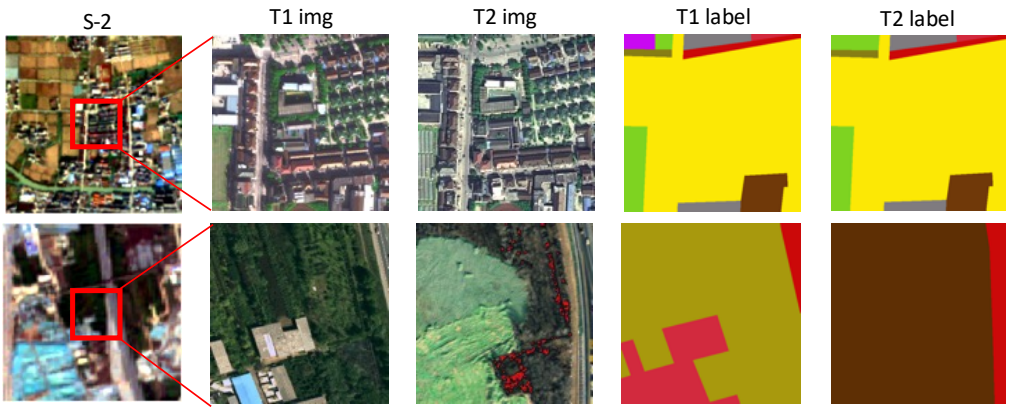

Figure 6: The alignment of Google Earth images and Sentinel images.

Table 11: Training and testing data split.

Data	Train	Test	Val
Complete	43,927	12,550	6,275
Xi'an	25,303	7,205	3,660
Jiaxing	18,624	5,345	2,615
Change Detection	16,998	4,813	2,413

dataset by capturing a wider range of features and details, providing multiple ways of observations on different kinds of human activities.

A.4 License

Use of the Google Earth images must respect the "Google Earth" terms of use. All images and their associated annotations in FUSU can be used for academic purposes only, and any commercial use is prohibited (CC BY-NC-SA 4.0).

A.5 Extra Experiment Results

A.5.1 Dataset Split

We show our training and testing dataset split in the URL link. In general, table 11 shows the details. Note that for change detection, we only select the patches that have changed pixels.

A.5.2 Implementation Details

For segmentation, we use a Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9 and a weight decay of 1e-4. The learning rate is 0.01, and a 'poly' scheduler with power 0.9 is applied. The batch size is 8 and the max training iterations are $80k$. For semantic change detection, we use AdamW as our optimizer and β_1 is 0.5 and β_2 is 0.999. The learning rate is 3e-4 and linearly decays are applied to 0 until trained for max epochs. The batch size is 8 and the max training epochs are 200. For binary change detection, the learning rate is 0.001, and other settings are the same as the semantic change detection. For FUSU-Net, the settings are the same as other methods in different tasks. For domain adaptation, we adopt the architectures in [5] and keep the default settings. All experiments are conducted on 4 NVIDIA GTX 4090 GPUs with 25 GB memory.

A.5.3 FUSU-Net Supervision

Fig. 7 shows the supervision in FUSU-Net. There are four outputs of FUSU-Net, i.e., change prediction, T1 prediction, time-series prediction, T2 prediction, and three labels, i.e., change label,

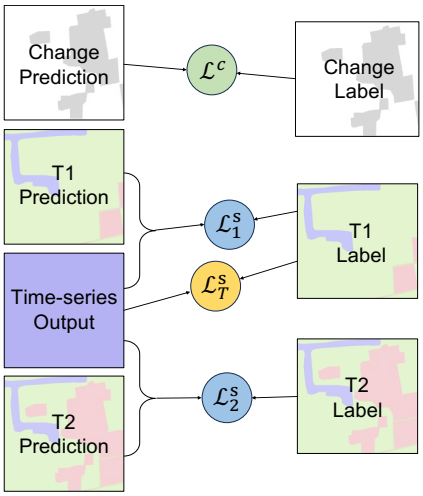

Figure 7: The visualization of supervision in FUSU-Net.

Table 12: Domain adaptation results obtained from training and testing on the whole dataset.

Domain	Method	IoU per class (%)																	mIoU
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	
Jiaxing ↓ Xi'an	FADA [38]	5.23	33.17	6.32	5.10	11.45	32.67	3.14	6.28	9.12	3.24	1.02	1.08	5.18	8.33	0.20	0.11	1.15	8.23
	PyCDA [39]	3.08	32.54	8.11	1.28	12.35	26.48	8.27	6.17	10.33	5.42	1.10	1.22	9.18	3.21	1.12	1.08	1.16	8.34
	CLAN [40]	4.33	24.67	6.42	3.15	18.34	39.78	4.22	8.17	14.23	4.14	2.08	3.18	3.11	6.33	1.16	1.41	1.57	9.63
	CBST [41]	17.21	10.43	5.34	2.11	28.78	39.32	2.07	5.38	7.22	3.42	7.12	2.03	2.14	2.21	0.44	0.45	1.13	9.24
	AdaptSeg [42]	7.23	35.78	14.54	3.12	13.67	25.22	3.18	6.34	8.28	5.13	2.11	3.12	3.14	2.25	1.15	1.19	1.24	9.68
	IAST [43]	8.13	36.42	15.78	4.12	15.89	27.65	5.12	7.42	10.21	6.38	3.21	4.19	4.33	8.45	2.81	2.22	2.35	10.93
Xi'an ↓ Jiaxing	FADA [38]	11.38	31.67	10.21	4.18	11.24	22.53	9.12	10.28	7.18	2.14	1.11	1.22	3.23	1.08	1.14	1.49	1.25	9.97
	PyCDA [39]	13.45	10.24	7.34	2.13	3.28	25.65	3.14	2.15	9.38	2.08	1.12	1.18	5.23	3.15	1.77	2.02	1.14	8.70
	CLAN [40]	11.38	13.42	12.24	4.28	10.53	17.68	4.17	8.28	13.42	2.14	1.11	1.25	5.21	1.08	0.78	1.34	1.21	9.24
	CBST [41]	6.42	23.78	20.12	2.11	28.89	22.34	3.21	14.27	17.45	2.14	1.23	1.28	3.14	1.16	0.62	1.39	1.18	10.21
	AdaptSeg [42]	6.28	40.18	6.11	2.17	18.42	40.78	2.13	6.42	4.11	2.18	1.21	1.27	1.34	1.16	1.24	1.29	1.35	10.63
	IAST [43]	8.38	38.27	8.12	3.24	16.42	38.68	3.24	7.24	4.28	3.15	1.18	1.25	2.27	2.17	2.23	2.28	2.32	10.89

T1 label and T2 label. To balance segmentation and change detection, we set the weight of change loss as 2, and the weights of segmentation losses as 1. The total loss function is calculated as:

$$\mathcal{L} = \mathcal{L}_1^s + \mathcal{L}_2^s + \mathcal{L}_T^s + 2 \times \mathcal{L}^c \tag{4}$$

A.5.4 Domain Adaptation

Because of the feature gaps between Jiaxing and Xi'an, as we discussed in Sec. 3 and Sec. 5.1, the FUSU dataset also has the ability to support domain adaptation. Here we evaluate the performance of 6 unsupervised domain adaptation methods on FUSU dataset, which include FADA [38], PyCDA [39], CLAN [40], CBST [41], AdaptSeg [42] and IAST [43]. IoU per class and mIoU over 17 classes are calculated.

Overall results. Table 12 shows the semantic change detection results. We can see there are some easy classes for all unsupervised domain adaptation methods (i.e., inland water-2, blank-6), which are similar to the results of semantic segmentation. Some classes bring challenges (i.e., storage-15,

Table 13: Top performances compared with other datasets.

Dataset	mIoU
GID [11]	90.79
ISPRS Potsdam[4]	82.17
LoveDA [5]	49.02
FLAIR [15]	54.51
FUSU	46.88

grass-17), indicating difficulty in adapting to feature changes in those specific categories. Also, we can see the interchange between the source domain and the target domain will affect the performance of domain adaptation tasks. Xi'an to Jiaxing task gets higher performance on blank-5 than the Jiaxing to Xi'an task. There isn't much disparity in performance between two mainstream approaches, i.e., adversarial training (AdaptSeg, CLAN, FADA) and self-training (CBST, PyCDA). In summary, these methods get unsatisfactory performance on our dataset. The results show little improvement compared to the source-only results, and in some cases, they are even worse. We hypothesize the following two reasons. First, general domain adaptation methods in the field of computer vision cannot adapt to the domain characteristics of the FUSU dataset, necessitating the development of improved methods. A customized method might achieve better results. Second, the categories in Jiaxing and Xi'an are discontinuous, with Jiaxing having more cropland and Xi'an having more urban buildings, resulting in a significant domain gap. This large gap makes it challenging for the methods to learn effectively.

A.5.5 Comparison with Other Datasets

We investigate the difficulty of several open-source segmentation datasets by comparing the top performances on these datasets via HRNet. We can see from Table 13 that FUSU has the lowest performance among all datasets, which indicates the difficulty of the FUSU dataset. We summarize the challenges of FUSU from two aspects. First, the feature gaps between Jiaxing and Xi'an increase the difficulty of this dataset. The training model must adapt with two main features during one end-to-end period. Second, the land use classification involves many understandings of human activities and production rather than land cover, which can directly correspond to pixel values. Therefore, relying solely on remote sensing imagery makes achieving high accuracy challenging. Multi-source data is needed for better performance.

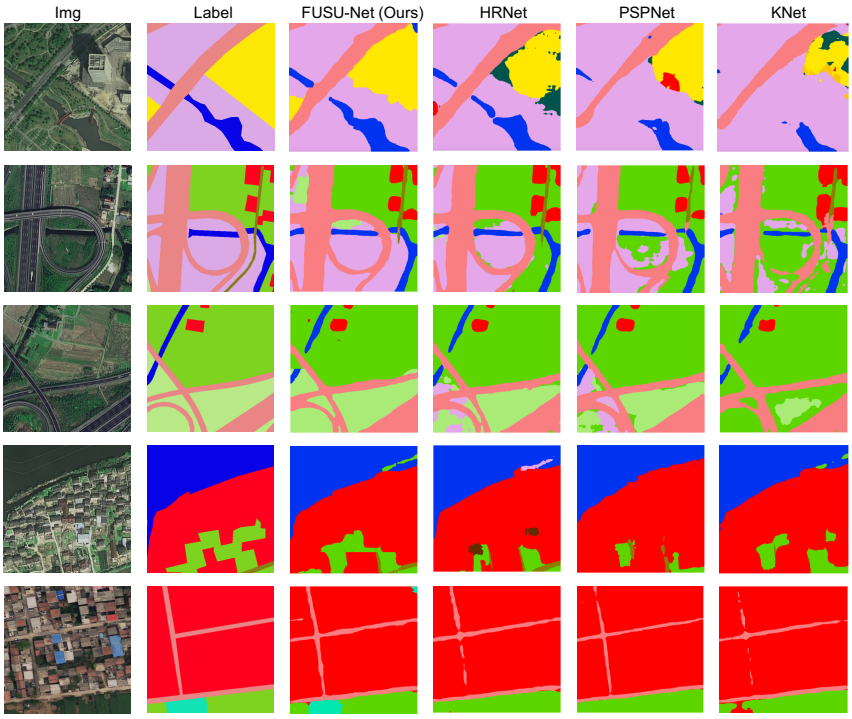

Figure 8: The visualization of results.

B Datasheet for FUSU

B.1 Motivation

- **Q1: For what purpose was the dataset created?** Was there a specific task in mind? Was there a particular gap that needed to be filled? Please provide a description.
  - The FUSU dataset was created to fill the gap in fine-grained land use data for urban areas. It aims to enable a deeper understanding of human activities and their impacts on the environment by providing detailed pixel-wise land use annotations and high-resolution spatial-temporal data. This dataset supports the development and benchmarking of deep learning models for tasks such as change detection and segmentation in urban environments, thereby enhancing contextual insights into urbanization and human activities.
- **Q2: Who created the dataset?**
  - This dataset is created by the Department of Earth System Science, Tsinghua University, and Tsinghua University (Department of Earth System Science)- Xi'an Institute of Surveying and Mapping Joint Research Center for Next-Generation Smart Mapping.
- **Q3: Who funded the creation of the dataset?**
  - The FUSU dataset creation is funded by the National Natural Science Foundation of China under Grant No. T2125006.
- **Q4: Any other comments?**
  - [N/A]

B.2 Composition

- **Q5: What do the instances that comprise the dataset represent?**
  - We provide bi-temporal high-resolution images with pixel-wise land use segmentation annotations along with corresponding Sentinel-1 and Sentinel-2 images center-around the high-resolution patches. This dataset spans over 847 km^2, covering 2 cities (Jiaxing and Xi'an) and 5 urban districts (Beilin, Yanta, Xincheng, Lianhu, Xiuzhou) in the southern and northern of China.
- **Q6: How many instances are there in total?**
  - We provide 62,752 image patches. Each patch contains 2 high-resolution images with a resolution of 0.2-0.5 m, 2 pixel-wise annotations with a resolution of 0.2-0.5 m, and 25 Sentinel images with a resolution of 10 m. The whole FUSU dataset comprises over 30 billion annotated pixels with 17 land use classes.
- **Q7: Does the dataset contain all possible instances or is it a sample?**
  - This dataset contains all possible instances in the selected study areas. We hope the community can contribute more instances in other places (e.g., other cities in China or other countries) with us.
- **Q8: What data does each instance consist of?**
  - Each patch contains 2 high-resolution images with a resolution of 0.2-0.5 m, 2 pixel-wise annotations with a resolution of 0.2-0.5 m, and 25 Sentinel images with a resolution of 10 m. The high-resolution image is with a size of 512×512 and R/G/B bands. The Sentinel image is with a size of 128×128 and 14 bands.
- **Q9: Is there a label or target associated with each instance?**
  - Yes, we provide bi-temporal pixel-wise labels with 17 land use classes and over 30 billion annotation pixels.
- **Q10: Is any information missing from individual instances?**
  - [No]
- **Q11: Are relationships between individual instances made explicit?**
  - [No]

- **Q12: Are there recommended data splits?**
  - Yes, we provide data splits in the GitHub link for reproducing.
- **Q13: Are there any errors, sources of noise, or redundancies in the dataset?**
  - Despite the rigorous quality control applied during visual interpretation, some errors are inevitable, particularly for classes that are visually difficult to distinguish. To minimize these errors, internal quality control involving multiple annotations has been conducted.
- **Q14: Is the dataset self-contained, or does it link to or otherwise rely on external resources?**
  - This dataset will be stored and distributed on the data website of Peng Cheng Laboratory. The link on the GitHub page will redirect to the data website.
- **Q15: Does the dataset contain data that might be considered confidential?**
  - [No]
- **Q16: Does the dataset contain data that might be offensive, insulting, or threatening?**
  - [No]
- **Q17: Does the dataset relate to people?**
  - The dataset may feature pedestrian or individuals, but the resolution of 20cm-50cm/pixel and the aerial perspective is not sufficient to recognize them uniquely.
- **Q18: Does the dataset identify any subpopulations?**
  - [No]
- **Q19: Is it possible to identify individuals from the dataset?**
  - [No]
- **Q20: Does the dataset contain sensitive data?**
  - [No]
- **Q21: Any other comments?**
  - [No]

B.3 Collection Process

- **Q22: How was the data associated with each instance acquired?**
  - The high-resolution images are collected from Google Earth. This platform pre-processes the images including correction, mosaic conduction, and so on. The labels are annotated and cross-checked by two expert teams manually. The Sentinel-2 and Sentinel-1 images are collected from Google Earth Engine platform with all necessary processing steps. All data is mapped onto the WGS84 coordinate reference system.
- **Q23: What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?**
  - We use the Google Earth API to collect all image data, and hire two expert teams to annotate the images.
- **Q24: If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**
  - [N/A]
- **Q25: Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**
  - We contracted geography experts from the remote sensing institute and university through a public call for tender to annotate the dataset. The creation of the dataset was facilitated by researchers and developers employed by us under their work contracts.
- **Q26: Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)?**

  - The collection of the data spans from Aug, 2018 to Aug, 2020. The high-resolution images are collected in Aug, 2018 and Aug, 2020. The Sentinel images are collected each month during this period. This dataset was created in 2022 after the processing steps and fine-tuned in 2023.
- **Q27: Were any ethical review processes conducted (e.g., by an institutional review board)?**
  - [No]
- **Q28: Does the dataset relate to people?**
  - [No]
- **Q29: Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**
  - [N/A]
- **Q30: Were the individuals in question notifies about the data collection?**
  - [N/A]
- **Q31: Did the individuals in question consent to the collection and use of their data?**
  - [N/A]
- **Q32: If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?**
  - [N/A]
- **Q33: Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?**
  - [No]
- **Q34: Any other comments?**
  - [No]

B.4 Preprocessing, Cleaning, and/or Labeling

- **Q35: Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?**
  - [No]
- **Q36: Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.**
  - [No]
- **Q37: Is the software used to preprocess/clean/label the instances available?**
  - [No]
- **Q38: Any other comments?**
  - [No]

B.5 Uses

- **Q39: Has the dataset been used for any tasks already?**
  - [No]
- **Q40: Is there a repository that links to any or all papers or systems that use the dataset?**
  - [No]
- **Q41: What (other) tasks could the dataset be used for?**

  - Actually FUSU dataset can support many other tasks because of its special attributes. First, because there exist different vision modalities in FUSU, we look forward to utilizing FUSU to explore multimodal fusion methods. Second, FUSU also offers the opportunities for super-resolution tasks.
- **Q42: Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?**
  - During the data creation, we use a data-expanding paradigm to enhance FUSU's temporal resolution, capturing more detailed changes during time series. We have discussed it in the Appendix. It is versatile enough to be extended to other readily available change detection datasets. We will provide the process steps and code accordingly for the community.
- **Q43: Are there tasks for which the dataset should not be used?**
  - [No]
- **Q44: Any other comments?**
  - [No]

B.6 Distribution

- **Q45: Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?**
  - Yes, the dataset will be open-source.
- **Q46: How will the dataset be distributed (e.g., tarball on website, API, GitHub)?**
  - The data will be available on GitHub, which redirects to the data website of Peng Cheng Laboratory.
- **Q47: When will the dataset be distributed?**
  - The demo will be released in June 2024, and the whole dataset, along with the dataset split, will be released in early October 2024.
- **Q48: Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?**
  - Yes, use of the Google Earth images must respect the "Google Earth" terms of use. All images and their associated annotations in FUSU can be used for academic purposes only, and any commercial use is prohibited (CC BY-NC-SA 4.0).
- **Q49: Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**
  - [No]
- **Q50: Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**
  - [No]
- **Q51: Any other comments?**
  - [No]

B.7 Maintenance

- **Q52: Who will be supporting/hosting/maintaining the dataset?**
  - The authors of this paper will support and maintain this dataset.
- **Q53: How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**
  - *shuai914@connect.hku.hk*
- **Q54: Is there an erratum?**
  - [No]

- **Q55: Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?**
  - Yes, this dataset will be updated if more data sources are available (e.g., DEM data, Night-time light data).
- **Q56: If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?**
  - [N/A]
- **Q57: Will older versions of the dataset continue to be supported/hosted/maintained?**
  - Yes, we will continue to provide support for the FUSU dataset.
- **Q58: If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?**
  - Communities can contact the hosts for correction or contributions to the FUSU dataset. All kind suggestions, extensions, augmentation, and contributions are welcome with the consideration of license and maintenance requirements.
- **Q59: Any other comments?**
  - [No]