FLOW-CDNET: A NOVEL NETWORK FOR DETECTING BOTH SLOW AND FAST CHANGES IN BITEMPORAL IMAGES

Haoxuan Li*
School of Computer Science and Technology
Northwestern Polytechnical University
Xi'an 710072, China
li_haoxuan@mail.nwpu.edu.cn

Chenxu Wei*
School of Computer Science and Technology
Northwestern Polytechnical University
Xi'an 710072, China
weichenxu@mail.nwpu.edu.cn

Haodong Wang
School of Computer Science and Technology
Northwestern Polytechnical University
Xi'an 710072, China
traslauc@mail.nwpu.edu.cn

Xiaomeng Hu
School of Computer Science and Technology
Northwestern Polytechnical University
Xi'an 710072, China
hxm2886@mail.nwpu.edu.cn

Boyuan An
School of Computer Science and Technology
Northwestern Polytechnical University
Xi'an 710072, China
741099841@mail.nwpu.edu.cn

Lingyan Ran
School of Computer Science and Technology
Northwestern Polytechnical University
Xi'an 710072, China
lran@nwpu.edu.cn

Baosen Zhang
Yellow River Institute of Hydraulic Research
Zhengzhou 450003, China
976129493@qq.com

Jin Jin
Yellow River Institute of Hydraulic Research
Zhengzhou 450003, China
jinjin@hky.yrcc.gov.cn

Omirzhan Taukebayev
Al-Farabi Kazakh National University
Almaty 050040, Republic of Kazakhstan
omirzhan.taukebayev@kaznu.edu.kz

Amirkhan Temirbayev
Al-Farabi Kazakh National University
Almaty 050040, Republic of Kazakhstan
amirkhan.temirbayev@kaznu.edu.kz

Junrui Liu†
School of Computer Science and Technology
Northwestern Polytechnical University
Xi'an 710072, China
liu.junrui@nwpu.edu.cn

Xiuwei Zhang
School of Computer Science and Technology
Northwestern Polytechnical University
Xi'an 710072, China
xwzhang@nwpu.edu.cn

July 4, 2025

ABSTRACT

Change detection typically involves identifying regions with changes between bitemporal images taken at the same location. Besides significant changes, slow changes in bitemporal images are also important in real-life scenarios. For instance, weak changes often serve as precursors to major hazards in scenarios like slopes, dams, and tailings ponds. Therefore, designing a change detection

*These authors contributed equally to this work.
†Corresponding author: liu.junrui@nwpu.edu.cn

network that simultaneously detects slow and fast changes presents a novel challenge. In this paper, to address this challenge, we propose a change detection network named Flow-CDNet, consisting of two branches: optical flow branch and binary change detection branch. The first branch utilizes a pyramid structure to extract displacement changes at multiple scales. The second one combines a ResNet-based network with the optical flow branch's output to generate fast change outputs. Subsequently, to supervise and evaluate this new change detection framework, a self-built change detection dataset Flow-Change, a loss function combining binary tversky loss and L2 norm loss, along with a new evaluation metric called FEPE are designed. Quantitative experiments conducted on Flow-Change dataset demonstrated that our approach outperforms the existing methods. Furthermore, ablation experiments verified that the two branches can promote each other to enhance the detection performance.

1 Introduction

In real-world monitoring scenarios, both **slow** and **fast** changes frequently coexist across various environments. For example, in applications such as slope monitoring, dam safety assessment, and tailings pond management, minor displacements of soil or ore blocks often reflect *slow changes*, while sudden collapses or structural failures correspond to *fast changes*. These phenomena may appear sequentially or simultaneously, thus posing significant challenges for accurate and robust change detection.

Traditional change detection (CD) techniques typically focus on identifying fast changes, where an object appears or disappears completely between two temporal images. These are often tackled using deep neural networks (DNNs) based on semantic segmentation or classification. In contrast, the detection of slow changes—characterized by partial displacement of objects over time—relies on optical flow estimation, which computes pixel-level correspondences between image pairs in the form of dense 2D displacement fields.

In this work, we define the concepts of slow and fast changes more formally: when an object is present in both images but changes its position or shape, we classify it as a *slow change*; if the object exists only in one of the bitemporal images, it is categorized as a *fast change*. This definition is consistent with real-world scenarios and reflects the need for a unified framework capable of simultaneously addressing both change types.

Extensive research has been conducted on the challenges of detecting either slow or fast changes. For slow changes, researchers commonly employ optical flow detection, since it aims to determine the pixel-wise correspondences between source and target images in the form of a 2D displacement field, allowing it to capture minor variations effectively. SpyNet [1] employs a coarse-to-fine approach that combines traditional methods with deep learning techniques. ContinuousFlow [2] combines occlusion and cost volume techniques with optical flow estimation. MaskFlownet [3] introduces a learnable occlusion mask in the asymmetric feature matching module, thus improving the effectiveness of optical flow prediction. LiteFlowNet [4] uses traditional brightness inconsistency mapping to address occlusion issues. RAFT [5] extracts per-pixel features and builds multi-scale 4D cost volumes for all pixel pairs. It maintains and updates a single fixed-resolution optical flow image with high resolution, so it has strong cross-dataset generalization ability. AccFlow [6] accumulates frame-to-frame optical flow to obtain long-distance cross-frame optical flow, adapting to optical flow estimation algorithms for arbitrary frame pairs. VideoFlow [7] can thoroughly explore and utilize multi-frame data, significantly enhancing the performance of optical flow estimation. For fast changes, researchers commonly employ change detection with deep neural networks. PSPNet [8] is the first to introduce the concept of pyramid pooling modules and integrate global contextual information. Due to its ability to leverage global contextual information through context aggregation from different regions, it has become a baseline method in the realm of deep change detection. Chen et al. [9] introduces a network model called DASNet for high-resolution image change detection. Liu et al. [10] utilizes semantic segmentation as an auxiliary task to aid change detection. This approach helps to learn more distinctive object-level features, thereby enhancing the quality of change detection results.

Despite recent advancements, there remains a critical limitation: existing methods cannot effectively detect slow and fast changes simultaneously. Optical flow models struggle with occlusions and object disappearance, while conventional CD networks are insufficiently sensitive to subtle, continuous transformations.

To overcome this limitation, we propose a novel framework named **Flow-CDNet**. This dual-branch architecture integrates optical flow estimation and change detection in a unified learning paradigm. Specifically, one branch estimates dense motion maps using a pyramid-based optical flow module, while the other branch performs binary change detection by leveraging both the original bitemporal images and the motion features. The joint learning mechanism enables mutual enhancement between the two branches, improving the network's capability to detect diverse change types with higher accuracy.

To facilitate model training and evaluation, we construct a dedicated dataset named **Flow-Change**, which includes synthetic bitemporal image pairs with annotated labels for both optical flow and binary change detection. Additionally, we design a composite loss function that jointly optimizes optical flow regression and change detection objectives. Furthermore, we introduce a new evaluation metric called **FEPE** (F1-score over End-Point Error), which provides a unified assessment of model performance across both slow and fast changes. Finally, we validate the generalization capability of Flow-CDNet on real-world dam bank images, demonstrating its effectiveness in practical applications.

The main contributions of this work are summarized as follows:

- We propose **Flow-CDNet**, a unified change detection framework that combines pyramid-based optical flow estimation with residual convolutional change classification in a dual-branch architecture, enabling simultaneous detection of both slow and fast changes in bitemporal imagery.
- We construct a new dataset named **Flow-Change** and propose a **composite loss function** and a novel evaluation metric (**FEPE**) to support end-to-end training and performance assessment.
- We assess the proposed approach on both real-world dam bank monitoring scenarios and the constructed synthetic dataset, where it effectively identifies both abrupt collapses and gradual deformations. These results, visualized in detail, not only demonstrate strong practical applicability but also highlight the complementary strengths of the dual-branch architecture through ablation studies.

2 Related Works

2.1 Optical Flow Estimation

Optical flow estimation techniques are broadly categorized into traditional and deep learning-based methods. Conventional approaches such as EpicFlow [11], DeepFlow [12], and MirrorFlow [13] propose various strategies to enhance estimation accuracy, yet they generally fall short in meeting real-time performance requirements.

The advent of deep learning revolutionizes the field. FlowNet [14] pioneers the use of convolutional neural networks for optical flow, achieving real-time operation but with limited precision. FlowNet2.0 [15] addresses this shortcoming by introducing a stacked architecture, utilizing synthetic datasets, and adopting improved training protocols, leading to notable accuracy gains. SpyNet [1] introduces a spatial pyramid structure with coarse-to-fine warping and residual prediction, enabling efficient large displacement flow estimation with low computational cost. LiteFlowNet3 [16] further refines performance by handling outliers in the cost volume through adaptive modulation prior to decoding and correcting distorted flows using nearby reliable estimates.

To better handle large displacement motion, GMFlowNet [17] and GMFlow [18] reformulate optical flow as a matching problem rather than a regression one, enhancing robustness and accuracy. Jiang et al. [19] show that precise flow estimation is attainable even when matching only a small subset of pixels, emphasizing the efficiency of sparse correspondences.

Transformer-based architectures also emerge as a powerful paradigm. FlowFormer [20] introduces transformers into this domain, employing a cost volume encoder for compact representation and a recursive decoder that iteratively refines flow using dynamic location queries. SAMFlow [21] extends this by integrating a frozen SAM image encoder and improving semantic perception of objects within the scene.

Robustness across scales and detail preservation are addressed by approaches like AnyFlow [22], which estimates flow from images of varying resolutions, excelling at capturing fine-grained motion. DistractFlow [23] contributes a novel data augmentation method by blending optical flow estimates with structurally similar disturbance images, aligning visual perturbations with real-world appearances.

In pursuit of greater efficiency, RAPIDFlow [24] incorporates NeXt1D convolutional blocks within a fully recursive feature pyramid, reducing computational load without compromising accuracy. MaxFlow [25] adopts 1D matching along with MaxViT transformers to significantly lower complexity while retaining strong performance. MatchFlow [26] enhances generalization through geometric pretraining and employs a QuadTree attention mechanism to boost adaptability across datasets.

Innovative attention-based designs continue to improve correlation modeling. CRAFT [27] revitalizes the correlation volume using a cross-attentional transformer, achieving resilience against blur and substantial motion. KPA-Flow [28] introduces kernel patch attention to strengthen local context modeling, setting new performance records on benchmarks like Sintel and KITTI.

Targeted solutions address specific limitations. I-RAFT [29] replaces zero-initialized flow with a multi-scale initialization strategy, improving accuracy while reducing model size. FlowDiffuser [30] reconceptualizes flow as a conditional generation problem using diffusion models, leveraging a noise-to-flow mechanism with a Conditional Recurrent Denoising Decoder. DeepPyNet [31] delivers a lightweight feature pyramid and a 4D correlation structure, achieving efficient performance with a minimal parameter footprint. PatchFlow [32] enables high-resolution flow estimation on resource-constrained devices through a two-stage patch-based method.

Novel theoretical frameworks also improve the modeling of optical flow. DEQFlow [33] formulates optical flow estimation as an equilibrium computation, where the solution corresponds to the fixed point of an implicit iterative process modeled by a deep equilibrium network. This formulation enables constant memory usage during training and inference, while also enhancing convergence stability. Equivariant Flow [34] introduces equivariant neural architectures to mitigate the inherent direction-dependent biases in conventional models, thereby improving generalization and robustness under varying motion dynamics.

Recent developments in remote sensing have introduced change detection networks that share foundational principles with optical flow, such as pixel-wise semantic differentiation and temporal consistency. Zhang et al. [35] propose a semi-supervised contrastive learning approach for semantic segmentation to enhance change detection, emphasizing effective feature discrimination even with limited labeled data. DifUnet++ [36] combines UNet++ with a differential pyramid module to better capture hierarchical spatial differences in satellite imagery. ADHR-CDNet [37] introduces attentive mechanisms to refine high-resolution spatial features, enabling precise detection of localized changes. These models illustrate the efficacy of segmentation-driven feature extraction, which complements optical flow estimation in temporally varying scenes.

Among all deep learning methods, RAFT [5] stands out as a foundational model that redefined the optical flow landscape through its iterative refinement strategy and all-pairs correlation volume. RAFT achieves a rare balance between accuracy and generalization by jointly encoding local and global motion cues while maintaining computational feasibility. Its elegant architecture and impressive benchmark performance have made it a cornerstone for many subsequent innovations, and it serves as the backbone for our proposed work.

2.2 Change Detection

Before the advent of deep learning, traditional change detection techniques play a pivotal role in remote sensing and computer vision. Among these, Principal Component Analysis (PCA) and Change Vector Analysis (CVA) are extensively employed in multispectral and hyperspectral imagery to reduce data dimensionality and accentuate spectral variations indicative of change. Post-classification comparison methods, which involve independently classifying each temporal image and subsequently contrasting the outputs, are widely utilized for thematic change mapping. Additionally, unsupervised clustering approaches such as K-Means and ISODATA are applied to difference images, enabling the identification of change regions without reliance on labeled data.

In recent years, the rapid progression of deep learning has revolutionized the landscape of change detection, delivering significant improvements in both accuracy and efficiency [38] [39]. Long et al. [40] pioneer the use of Fully Convolutional Networks (FCNs) for end-to-end semantic segmentation, laying the groundwork for numerous deep learning-based change detection frameworks. Building upon this paradigm, Dault et al. [41] propose three FCN-based architectures tailored to change detection tasks. Tezcan et al. [42] introduce BSUVNet, a supervised model leveraging FCNs for occlusion-aware background subtraction in video sequences, with auxiliary semantic segmentation inputs to enhance detection fidelity.

To address the challenges of dynamic scene analysis, Ou et al. [43] present the Deep Frame Difference Convolutional Neural Network (DFDCNN), which incorporates two specialized subnetworks—DifferenceNet and AppearanceNet—that collaboratively predict foreground segmentation maps. Cheng et al. [44] develop a multi-flow convolutional framework for action recognition in video data, employing sparse temporal sampling, frame differencing, and multi-branch learning to extract rich spatiotemporal features. Zheng et al. [45] propose a hybrid architecture combining convolutional layers with transformer modules, designed to extract both global and local change features through a tri-branch structure enhanced by spatial and channel-level dual attention mechanisms. Parallel research efforts also target heterogeneous change detection problems [46].

Yang et al. [47] introduce DLCDet, a dictionary learning-based framework integrated with a feature pyramid network and dual supervision strategy, effectively addressing semantic inconsistencies across temporal scenes and improving detection under seasonal variation. Zhang et al. [48] propose a deep Siamese network augmented with a contextual transformer module, employing multi-scale fusion and self-attention to strengthen bi-temporal feature alignment and pixel-wise precision. Sun et al. [49] devise a hybrid model combining Convolutional Neural Networks (CNNs) and

Graph Neural Networks (GNNs), enhanced with hierarchical feature supervision to facilitate fine-grained semantic interactions in high-resolution change detection.

Wang et al. [50] design a building change detection pipeline that synergizes pixel-level and object-level cues to generate saliency-guided difference maps, further refined through fuzzy clustering and a deep classification network. In low-data regimes, Paul et al. [51] explore a transfer learning strategy utilizing a pre-trained VGG19 backbone and a lightweight FCN, followed by SVM classification, yielding improved generalization. Nie et al. [52] develop a semi-supervised framework integrating Generative Adversarial Networks (GANs), residual Siamese networks, flow alignment, and atrous convolutions to robustly handle limited labeled data scenarios in remote sensing.

Zhu et al. [53] propose ChangeViT, a Vision Transformer-based architecture featuring a detail-capturing mechanism and feature injection module, achieving state-of-the-art performance on large-scale fine-grained benchmarks. Dong et al. [54] formulate a multimodal change detection approach that integrates image-text embeddings from CLIP, supplemented by a differential compensation module to enhance semantic change localization. Yin et al. [55] present a vision-language joint learning framework for box-supervised change detection, leveraging transformer-based fusion of visual and textual modalities.

Chen et al. [56] introduce SGANet, a geometry-aware Siamese network incorporating RGB imagery and monocular depth estimation, guided by cross-attention mechanisms to improve spatial localization through geometric cues. Meng et al. [57] propose ChangeAD, which combines bi-temporal alignment with differential feature integration to enhance robustness against seasonal and illumination variations. Finally, Fazry et al. [58] develop LocalCD, a locality-sensitive Vision Transformer that substitutes conventional feed-forward layers with depth-wise convolutions, enabling more accurate delineation of change boundaries.

3 Proposed Method

3.1 Network Architecture

In this section, we build up a framework named Flow-CDNet for simultaneous slow and fast change detection. As shown in Fig. 1, it consists of an optical flow detection branch (OFbranch) and a classical change detection branch (CDbranch). The OFbranch employs a pyramid structure to extract displacement changes at multiple scales, as depicted in Fig. 2. The CDbranch utilizes a network architecture based on spatial pyramid pooling to transform the output results into binary images with CNN, illustrated in Fig. 3. The following subsections provide details.

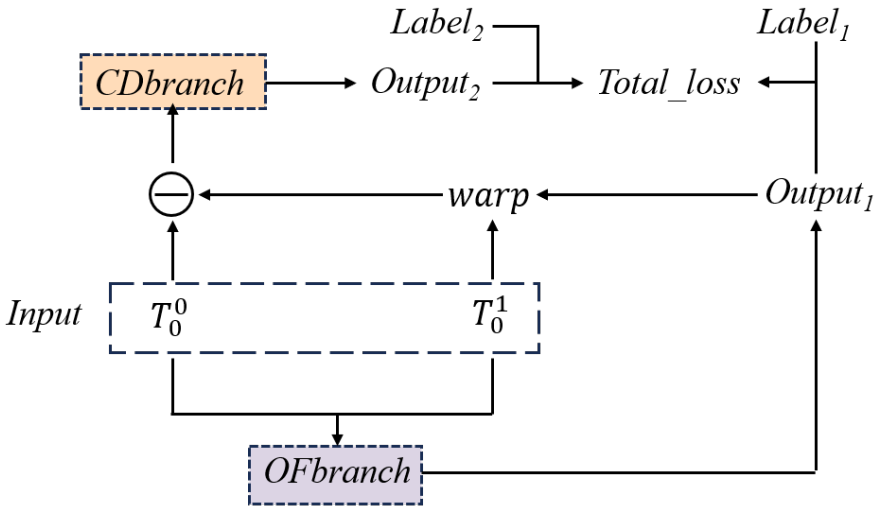

Figure 1: Overview of the proposed Flow-CDNet framework.

3.1.1 Structure of Flow-CDNet.

The proposed Flow-CDNet model is designed to jointly learn optical flow estimation and change detection from bitemporal image pairs, thereby enabling accurate identification of both fast and slow scene changes. Given a pair of temporally separated images T_0^0 and T_0^1, which represent the same scene captured at two different time points,

the network performs a sequence of processing steps to extract motion and change-related information through two interdependent branches.

Firstly, the input image pair is forwarded through the Optical Flow Estimation Branch, which is responsible for estimating the pixel-wise motion between T_0^0 and T_0^1. This branch outputs a dense motion estimation map, denoted as $output_1$, representing the displacement of each pixel from one image to the other.

Following this, the image T_0^1 is warped according to the estimated flow field $output_1$, resulting in a motion-aligned image. This operation aims to minimize the misalignment between the two images due to object or scene motion, thus allowing more accurate identification of residual differences attributable to genuine scene changes.

Subsequently, the absolute pixel-wise difference between the original image T_0^0 and the motion-compensated version of T_0^1 is computed. This difference map serves as a preliminary indicator of potential changes, though it may still contain noise or artifacts, especially in regions with inaccurate motion estimation.

To address this, the model incorporates an adaptive mask mechanism designed to emphasize regions likely to exhibit fast or abrupt changes. This mechanism adaptively weights the features based on motion uncertainty and intensity, thereby enhancing the network's ability to localize meaningful change patterns while suppressing false positives from low-confidence areas.

The refined difference features, together with the flow estimation $output_1$, are then fed into the Change Detection Branch. This second branch processes the combined information and produces a binary or probabilistic change map, denoted as $output_2$, which reflects the spatial distribution of detected changes between the two input frames.

By explicitly using $output_1$ as an auxiliary input to the CDbranch, the network effectively integrates both motion and appearance cues, forming a tightly coupled dual-branch architecture. This integrated design—central to the Flow-CDNet framework—has been empirically validated through ablation studies, which demonstrate that the inclusion of optical flow guidance significantly enhances the precision and robustness of change detection, particularly in the presence of dynamic background motion.

Finally, the overall training objective of the network is formulated as a multi-task loss function, which jointly optimizes both branches. The total loss aggregates the error terms from the flow estimation branch (related to $output_1$) and the change detection branch (related to $output_2$), thereby encouraging mutual reinforcement between the two tasks during end-to-end training.

3.1.2 Optical Flow Detection Branch.

As shown in Fig. 2, the input to the OFbranch is a pair of bitemporal images, i.e., T_0^0 and T_0^1 with the size of (H, W). The objective is to estimate a dense displacement field $\mathbf{f} = (f^1, f^2) \in \mathbb{R}^{H\times W\times 2}$, which maps each pixel (u, v) in T_0^0 to its corresponding coordinates $(u', v') = (u + f^1(u), v + f^2(v))$ in T_0^1. In our model, the OFbranch is implemented using the RAFT (Recurrent All-Pairs Field Transforms) architecture [5], which achieves state-of-the-art performance in optical flow estimation through iterative updates based on a high-resolution correlation volume and learned update operators. OFbranch consists of 3 main components, feature extraction, computing visual similarity and iterative updates.

For feature extraction, The feature encoder g_θ, implemented as a convolutional network with residual blocks, extracts high-dimensional features from both input images. It progressively downsamples the spatial resolution to $1/8$ of the original dimensions, generating feature maps $g_\theta(T_0^0)$ and $g_\theta(T_0^1) \in \mathbb{R}^{H/8\times W/8\times 256}$. A separate context encoder h_θ, sharing the same architecture as g_θ, processes only T_0^0 to produce a contextual feature map $h_\theta(T_0^0)$. This contextual information encodes semantic priors to guide motion boundary refinement during iterative updates.

For visual Similarity Computation, visual similarity between T_0^0 and T_0^1 is quantified by computing the inner product of all feature pairs across the two images. The resulting 4D correlation volume $C \in \mathbb{R}^{H/8\times W/8\times H/8\times W/8}$ is defined as:

$$C_{ijkl} = \sum_{h=1}^{256} g_\theta(T_0^0)_{ijh} \cdot g_\theta(T_0^1)_{klh}, \tag{1}$$

where (i, j) and (k, l) denote spatial positions in T_0^0 and T_0^1, respectively. To capture both large and small displacements, a multi-scale pyramid $\{C^1, C^2, C^3, C^4\}$ is constructed by applying average pooling to the last two dimensions of C. Pooling kernel sizes $\{1, 2, 4, 8\}$ correspond to progressively coarser resolutions, enabling hierarchical matching across varying displacement ranges.

For iterative updates, The optical flow field $\mathbf{f}$ is iteratively refined from an initial estimate $\mathbf{f}_0 = \mathbf{0}$ using a lightweight convolutional GRU unit. At each iteration k:

(a) For each pixel $\mathbf{x} = (u, v)$ in T_0^0, the current flow estimate $\mathbf{f}_k$ maps $\mathbf{x}$ to a correspondence $\mathbf{x}' = (u + f_k^1(u), v + f_k^2(v))$ in T_0^1. A local grid $\mathcal{N}(\mathbf{x}')_r = \{\mathbf{x}' + \mathbf{dx} \mid \|\mathbf{dx}\|_1 \leq r\}$ is defined around $\mathbf{x}'$, and multi-scale correlation features are retrieved via bilinear interpolation from the pyramid $\{C^1, C^2, C^3, C^4\}$.

(b) The retrieved correlation features, flow features (encoded from $\mathbf{f}_k$), and contextual features $h_\theta(T_0^0)$ are concatenated into a unified input tensor.

(c) The input tensor is processed by a convolutional GRU cell to predict a flow increment $\Delta\mathbf{f}$. The hidden state $\mathbf{h}_t$ and flow update are governed by:

$$
\begin{aligned}
\mathbf{z}_t &= \sigma\left(\text{Conv}_{3\times3}([\mathbf{h}_{t-1}, \mathbf{x}_t], \mathbf{W}_z)\right), \\
\mathbf{r}_t &= \sigma\left(\text{Conv}_{3\times3}([\mathbf{h}_{t-1}, \mathbf{x}_t], \mathbf{W}_r)\right), \\
\tilde{\mathbf{h}}_t &= \tanh\left(\text{Conv}_{3\times3}([\mathbf{r}_t \odot \mathbf{h}_{t-1}, \mathbf{x}_t], \mathbf{W}_h)\right), \\
\mathbf{h}_t &= (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \tilde{\mathbf{h}}_t,
\end{aligned}
\tag{2}
$$

where $\mathbf{x}_t$ denotes the fused features, and $\odot$ represents element-wise multiplication. The flow field is updated as $\mathbf{f}_{k+1} = \mathbf{f}_k + \Delta\mathbf{f}$. This recurrent process, with shared weights across iterations, enables stable convergence even after 100+ updates, avoiding error propagation inherent in coarse-to-fine approaches.

The final low-resolution flow field $\mathbf{f}_N \in \mathbb{R}^{H/8\times W/8\times 2}$ is upsampled to full resolution $output_1 \in \mathbb{R}^{H\times W\times 2}$ using a convex combination strategy.

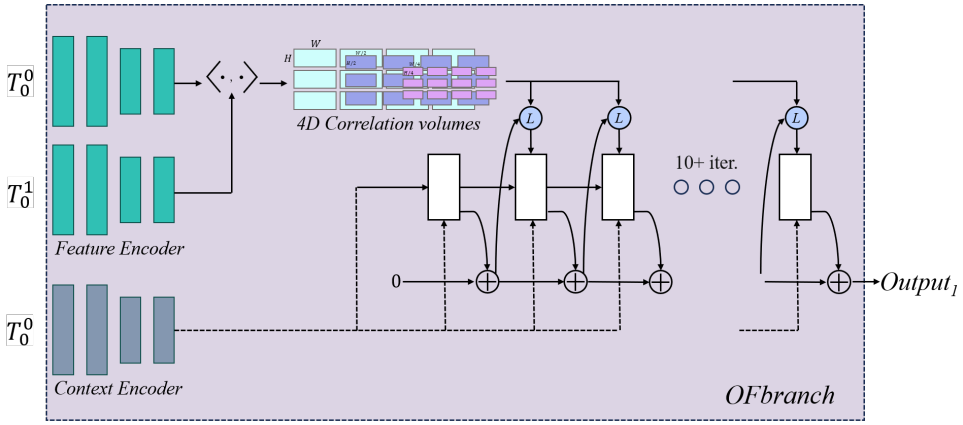

Figure 2: OFbranch.

3.1.3 Change Detection Branch.

As shown in Fig. 3, the CDbranch has three inputs: the input image T_0^0 and T_0^1, and $output_1$ obtained from the OFbranch. First, the image T_0^1 is warped using the predicted optical flow $output_1$ to generate a motion-compensated image, denoted as $w(T_0^1, output_1)$, then compute the absolute difference between T_0^0 and $w(T_0^1, output_1)$. Subsequently, a mask mechanism employing optical flow magnitude analysis dynamically identifies slow-changing regions, effectively suppressing their interference in the change detection process. Second, input the absolute difference result into the convolutional block using ResNet50 as backbone, and the output is denoted as F_0. Third, input F_0 into four parallel average pooling operations to generate four feature maps with different sizes, noted as F_1, F_2, F_3, F_4. Fourth, upsample these four feature maps to the same size as F_0, and perform channel stacking with the original feature map F_0, then pass the stacked feature maps through a 3×3 convolution, regularization and ReLU to obtain the feature map F_5 with 512 channels. Fifth, through 3×3 convolution and Sigmoid, a single-channel binary change feature map F_6 is obtained as the final output of the CDbranch, denoted as $output_2$.

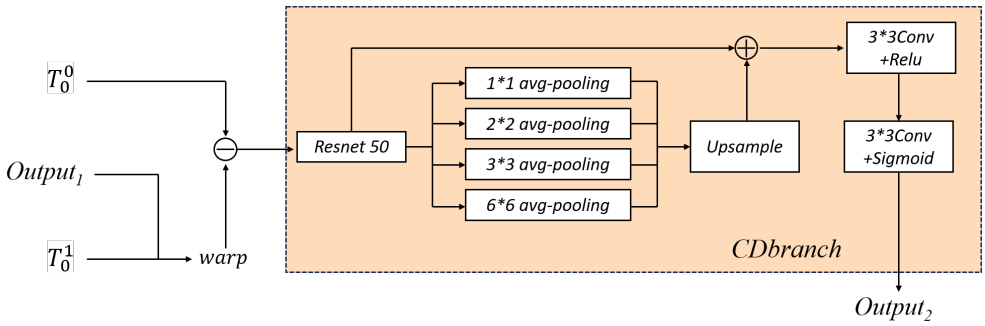

Figure 3: CDbranch.

3.2 Loss Function

To evaluate the results of optical flow detection, it is necessary to remove the binary changed regions ($label_2$) and extract only the regions containing slow change, then calculate the L2 norm loss with $label_1$ and $output_1$, as shown in Equation (3).

$$loss_{l2} = ||output_1 - label_1||_2 \cdot (1 - label_2) \tag{3}$$

To evaluate the results of change detection, we employ the change detection labels ($label_2$) and $output_2$ to compute the Tversky Loss[59], outlined in Equation (4).

$$loss_{Tversky} = \frac{masked\_gt}{masked\_gt + \alpha \cdot wrong\_classified + \beta \cdot unmasked\_gt} \tag{4}$$

Where $masked\_gt = output_2 \cdot label_2$, $unmasked\_gt = output_2 \cdot (1 - label_2)$, $wrong\_classified = (1 - output_2) \cdot label_2$, α and β are hyperparameters that control the penalty weights for $wrong\_classified$ and $unmasked\_gt$, respectively. By considering the distinct magnitudes of the two losses, we assign a weight ψ to Tversky loss. The overall training loss is presented in Equation (5).

$$loss_{total} = loss_{l2} + \psi \cdot loss_{Tversky} \tag{5}$$

3.3 Evaluation Metric

The widely used evaluation criterion, F1-score and EPE (End-Point Error) are adopted to evaluate the performance of binary change detection and optical flow. The F1-score is the statistical analysis of predictions, considering the value of true positive (TP), false positive (FP), and false negative (FN). They are defined as:

$$F_1 = 2/(Precision^{-1} + Recall^{-1}) \tag{6}$$

$$Precision = TP/(TP + FP) \tag{7}$$

$$Recall = TP/(TP + FN) \tag{8}$$

The EPE metric calculates the euclidean distance between the estimated optical flow($\vec{F}$) and the ground truth ($\vec{F_{gt}}$), as defined in Equation (9):

$$EPE = \sqrt{|\vec{F} - \vec{F_{gt}}|} \tag{9}$$

Considering regions with subtle motion, the mean EPE (mEPE) metric is adopted, which calculates the average error over all pixels within the union of regions with offset in the ground truth labels (denoted as Q) and regions with offset in the predicted outputs (denoted as Q^*), as described in Equation (10):

$$mEPE = \sum_{i \in Q \cup Q^*} \frac{\sqrt{|\vec{F_{gt}^i} - \vec{F^i}|}}{||Q \cup Q^*||} \tag{10}$$

While, to simultaneously evaluate binary change detection and optical flow results, we need a comprehensive evaluation indicator, so a new evaluation criterion namely FEPE is designed to combine the F1-score and the mEPE metric with ϵ being a low perturbation, as shown in Equation (11).

$$FEPE = \frac{F_1}{mEPE + \epsilon} \tag{11}$$

When F1-score becomes larger/smaller (indicating better/worse change detection performance) and mEPE metric becomes smaller/larger (indicating better/worse optical flow estimation performance), FEPE metric tends to become larger/smaller (indicating better/worse prediction performance). This design allows for a comprehensive evaluation that considers both change detection and optical flow estimation.

4 Experiments

4.1 Synthetic Flow-Change Dataset

Due to the relatively low occurrence of hazards such as deformations in dam bank and other slope areas, there is a scarcity of preserved image data, which makes it challenging to provide sufficient training data for deep models. Additionally, there is currently no existing change detection dataset that includes both fast and slow changes simultaneously. To train and evaluate the proposed method, a new synthetic dataset named Flow-Change is built by integrating FlyingChairs dataset [14] (an optical flow dataset) with PASCAL VOC 2007 dataset [60] (a semantic segmentation dataset).

As shown in Fig. 4, each data instance comprises a total of 4 images with size of 512×384 pixels, including a bitemporal image pair synthetized from FlyingChairs and PASCAL VOC, optical flow detection label map, and change detection label map. The training set consists of 11,736 pairs of images, while the test set includes 2,935 pairs of images.

The specific process for creating this dataset is as follows: each pair of bitemporal images is sequentially selected from the FlyingChairs dataset as the background. Then, potential change object regions are extracted from the PASCAL VOC 2007 dataset (e.g., humans, animals, vehicles), and subjected to random transformations including scaling, rotation, and random channel shuffle, then these objects are pasted onto the images to simulate fast change scenarios. Additionally, typical data augmentation techniques such as image brightness and contrast enhancement are applied to enrich the synthesized dataset. The corresponding optical flow detection labels from FlyingChairs and binary change detection labels, computed based on the pasting positions, are recorded as the detection ground truth.

4.2 Experiment on Flow-Change Dataset

Experiment settings. Our experiments are conducted on a GPU cluster comprising four NVIDIA GeForce GTX 4090 accelerators. The network architecture implements distinct learning rate configurations: the optical flow (OF) branch is initialized with 1e-5 learning rate, while the change detection (CD) branch employs 1e-4. We configure the optimization framework with AdamW and train for 1,000 epochs using batch size 4. We adopt the Tversky Loss for the CD branch, with α set to 0.7 and β set to 0.3. The multi-task loss weighting coefficient ϕ maintains a fixed ratio of 10 throughout all experiments.

Comparison experiments. The quantitative comparison results on Flow-Change dataset are displayed in Table 1. Since there is no similar network for both fast and slow change detection, different backbones of optical flow estimation are adopted to construct Flow-CDNet like networks. Flow-CDNet-L utilizes LiteFlowNet[4] and CDNet as backbone, Flow-CDNet-R utilizes RAFT[5] and CDNet as backbone, Flow-CDNet utilizes SpyNetC2 and CDNet as backbone. The best-performing results are shown in bold, and the second-best results are underlined. As shown in Table 1, it is evident that Flow-CDNet outperforms all compared backbones and achieves the highest FEPE metric of 0.869. Compared to the second-best optical flow detection backbone RAFT, our method achieves an EPE improvement of 1.382 and a 3.2% higher F1-score than the next best backbone (SpyNet+CDNet).

Ablation Study. Ablation experiments are conducted on Flow-Change dataset to verify the effectiveness of the OFbranch and the CDbranch, and further to analyze the impact of different OFbranch design on network performance.

Table 2 displays the ablation study results. Compared with only using the OFbranch, Flow-CDNet achieves a higher mEPE metric, indicating that the result ($output_2$) of change detection has a positive contribution to the overall detection results. In contrast, compared with only using the change detection branch, Flow-CDNet exhibits a higher F1-score. This indicates that the result ($output_1$) of optical flow detection also contributes to the overall detection results. This experiment demonstrates that the two branches can mutually enhance each other, improving Flow-CDNet's ability to detect both slow and fast changes.

Visualization. To qualitatively assess the performance of Flow-CDNet on synthetic scenarios with well-controlled dynamic object motion, we utilize visualizations from the Flow Dataset. As shown in Fig. 5, each row corresponds to a different scene containing foreground fast change objects and varied backgrounds. From left to right, the columns

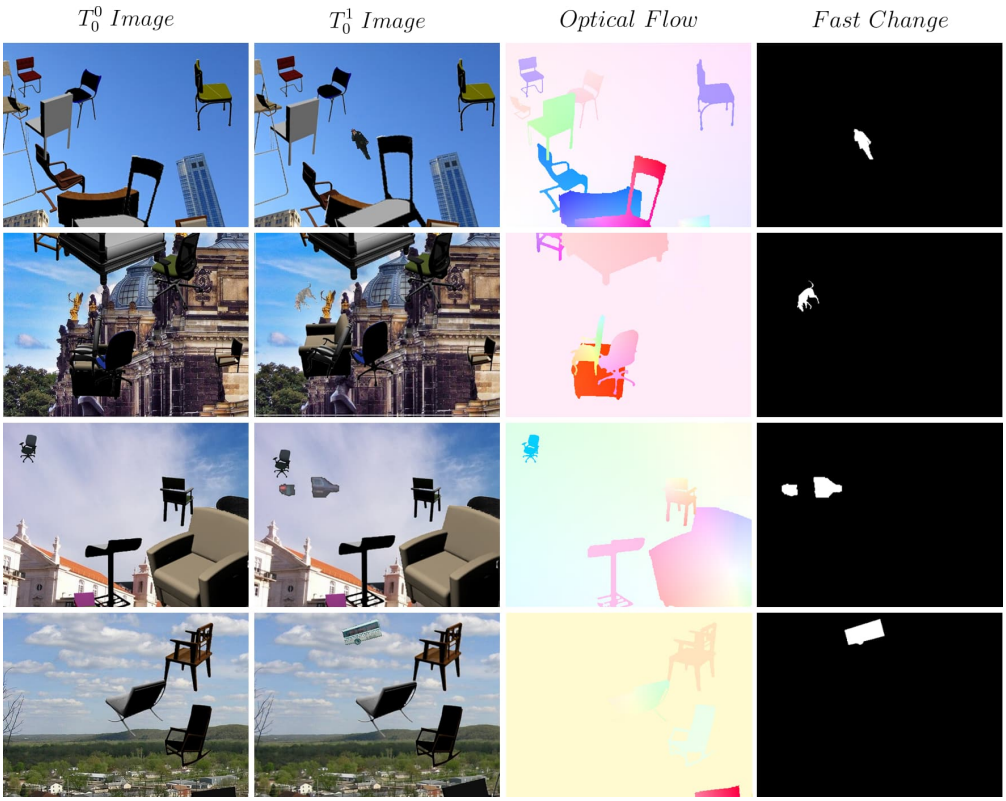

Figure 4: Dataset Visualization, T_0^0 represents image in the first frame, T_0^1 represents image in the second frame, Optical Flow column shows the ground truth transition between frame one and two, and Fast Change column shows ground truth of the item that only showed up in frame two.

Methods	F1-score↑	mEPE↓	FEPE↑
CDNet[8]	0.753	-	-
SpyNet[1]	-	3.383	-
LiteFlowNet[4]	-	6.433	-
RAFT[5]	-	<u>2.409</u>	-
Flow-CDNet-L(LiteFlowNet+CDNet)	0.821	5.720	0.144
Flow-CDNet-S(SpyNet+CDNet)	<u>0.860</u>	2.798	<u>0.308</u>
Flow-CDNet	**0.892**	**1.027**	**0.869**

Table 1: Quantitative comparison results with the state-of-the-art methods on the Flow-Change dataset. The best in bold, and the second-best is underlined.

display the bitemporal input images, the ground truth optical flow, and the binary change masks. In each case, the bitemporal images present noticeable object displacements between two time points, simulating real-world movement such as translation, rotation, or partial occlusion. The ground truth flow maps illustrate the pixel-wise motion fields, where distinct colors represent direction and magnitude of motion, facilitating fine-grained evaluation of optical flow estimation accuracy. Meanwhile, the ground truth change masks (final column) provide supervision for the change detection task by highlighting regions where significant structural changes occur. This visualization demonstrates our task of making the model's able to jointly learn both dense motion estimation and sparse change detection, which are essential for handling diverse temporal variations. The Flow Dataset thus serves as a valuable benchmark for validating Flow-CDNet's capacity to capture object-level dynamics under controlled conditions.

4.3 Experiment on Real-World Data

To further evaluate the generalization ability and real-world applicability of Flow-CDNet beyond synthetic benchmarks, we conduct experiments on a curated real-world test set composed of dam bank imagery. This dataset comprises 20

Branch		Metric		
OFbranch	CDbranch	F1-score↑	mEPE↓	FEPE↑
√	-	-	2.409	-
-	√	0.753	-	-
√	√	**0.892**	**1.027**	**0.869**

Table 2: Ablation study on selecting different branches. "√" indicates that the network utilizes the corresponding branch. The best-performing result is in bold.

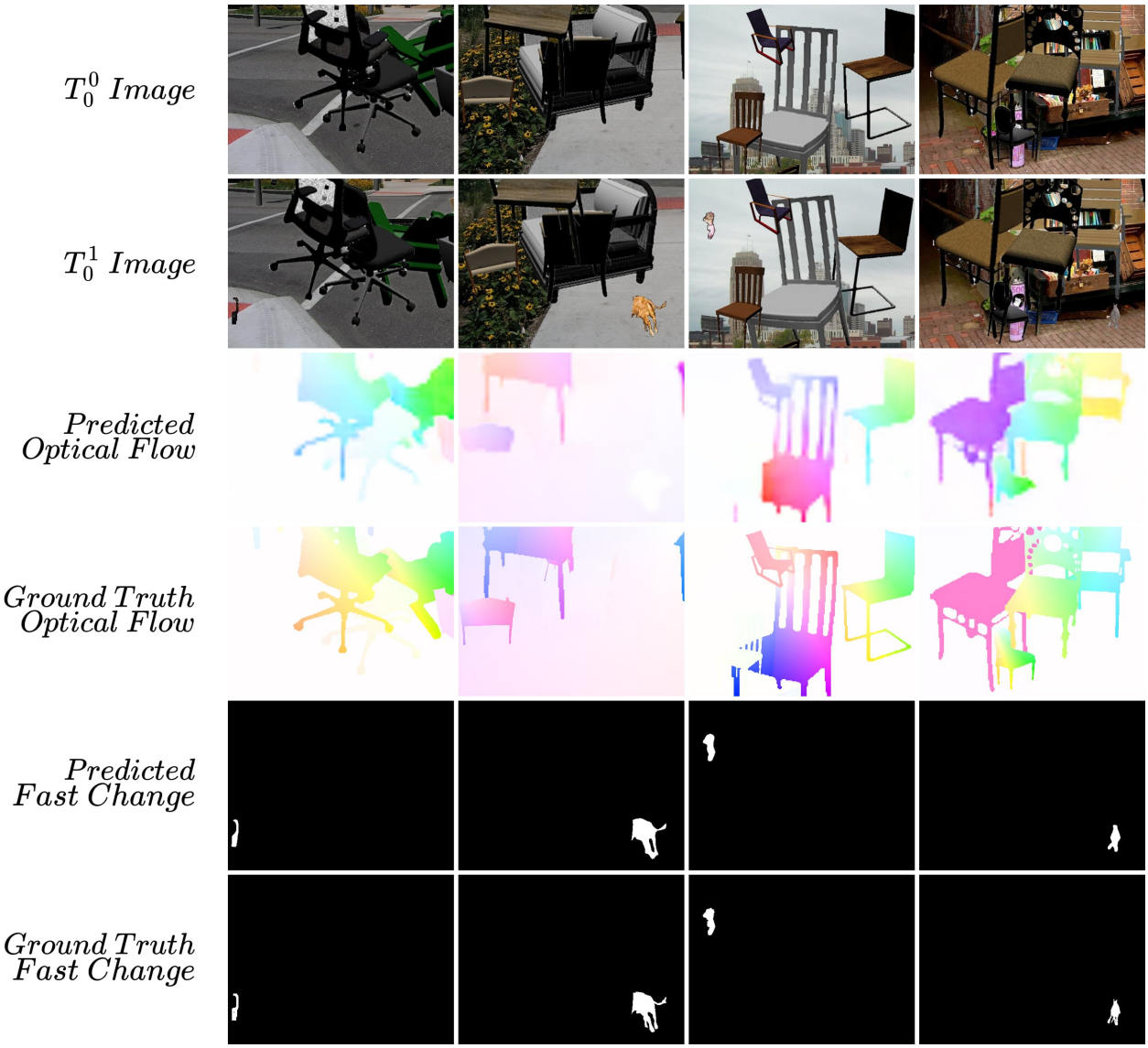

Figure 5: Model result visualization on the Synthetic Flow-Change Dataset

bitemporal image pairs acquired from 12 distinct dam sites, with each pair capturing the same location at two separate time points, denoted as T_0^0 and T_0^1. The imagery originates from two primary sources: ground-based fixed-position cameras (commonly referred to as trail cameras) and unmanned aerial vehicles (UAVs). The use of trail cameras represents a form of close-range remote sensing, offering continuous, localized monitoring, while UAV imagery provides broader spatial coverage with flexible vantage points. The selected sites reflect a diverse range of real-world change conditions, encompassing both abrupt structural events (e.g., sudden collapses or erosion) and gradual morphological

deformations over time. This multi-source dataset enables a comprehensive assessment of the model's robustness under varied sensing modalities and temporal change dynamics.

Ground truth (GT) annotations were manually constructed to identify regions exhibiting noticeable change between the two time points. However in real-world scenarios, the accurate labeling of fine-grained optical flow vectors presents significant challenges due to complex spatial displacements and the lack of precise reference data. To address this limitation, a unified annotation strategy is adopted, wherein all observable change regions are marked without categorizing them as abrupt or gradual. Consequently, both the detections from the optical flow estimation branch and the change detection branch are considered valid if they correctly localize these annotated regions.

The model tested in this setting is a pretrained Flow-CDNet, initially optimized on the Flow-Change dataset. During inference, each image pair is processed through the network without any domain-specific fine-tuning, thereby serving as a measure of its robustness to previously unseen data. Representative qualitative results are illustrated in Fig. 6. In columns one and two, examples of fast changes are shown, where catastrophic deformation results in substantial structural loss. These are clearly detected by the change detection branch, which effectively captures discrete region-level differences. Columns three to five present examples of slow, progressive surface deformation. Such subtler variations, which may be overlooked by direct difference-based methods, are more effectively revealed through the optical flow estimation branch that encodes temporal pixel-wise displacement information.

This complementary interplay between the two branches—one focusing on binary regional change and the other on continuous motion modeling—underscores the design rationale of Flow-CDNet. Despite the absence of separate GT for fast versus slow changes, the model demonstrates strong performance in capturing both types under a unified evaluation metric. After selecting appropriate change thresholds for different scenes, the final evaluation yields an F1-score of 0.8126, thereby quantitatively affirming the effectiveness of the proposed method. These findings validate the model's applicability to diverse real-world scenarios and confirm its capability to generalize to complex, uncontrolled environments where both discrete and continuous changes co-occur.

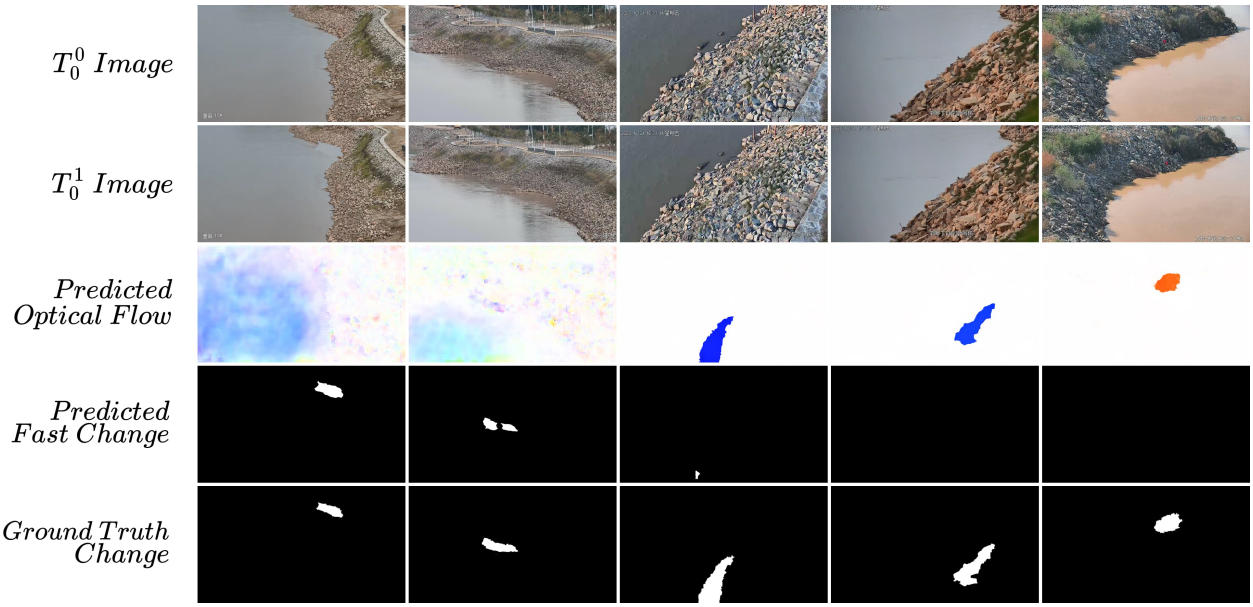

Figure 6: Visual comparison of model performance on real-world data. From top to bottom, the rows display: the input image at time T_0^0, the input image at time T_0^1, the output optical flow estimation from the proposed Flow-CDNet, the output fast change mask from the proposed Flow-CDNet, and the ground truth mask.

4.4 Model Efficiency Analysis

As shown in Table 3, we evaluate the computational efficiency of three variants of our proposed Flow-CDNet model. The standard Flow-CDNet achieves 14.08 FPS on an RTX 4090 and 12.50 FPS on an RTX 2080, with an inference time of 0.071 seconds and 0.080 seconds, respectively, and a computational cost of 401.8 GFLOPs. The lightweight variants, Flow-CDNet-L and Flow-CDNet-S, offer higher efficiency, reaching 21.73 FPS and 33.33 FPS on an RTX 4090, and 19.23 FPS and 25.32 FPS on an RTX 2080, with reduced computational costs of 215.3 GFLOPs and 212.6 GFLOPs, respectively.

In our downstream task scenario, current inference speed is sufficient to meet application requirements, as real-time performance is not a critical constraint. This allows us to leverage the RAFT-based Flow-CDNet, which prioritizes flow estimation accuracy over computational efficiency compared to lighter backbones like LiteFlowNet and SpyNet. For scenarios demanding higher real-time performance, deploying the model on more powerful hardware can further enhance inference speeds. The reported results across both GPU configurations demonstrate that all variants maintain practical efficiency, with the lightweight models providing significant speed improvements while preserving performance.

Table 3: Model Efficiency Comparison on Different GPUs

Model	FLOPs (G)	Inference Time (s)		FPS	
		RTX 4090	RTX 2080	RTX 4090	RTX 2080
Flow-CDNet	401.82	0.071	0.080	14.08	12.50
Flow-CDNet-L	215.30	0.046	0.052	21.73	19.23
Flow-CDNet-S	212.58	0.030	0.038	33.33	25.32

4.5 Analysis of Individual Branch Contributions

4.5.1 Comparison of Optical Flow Branch Variants

To ascertain the influence of the optical flow estimation backbone on the overall performance of the Flow-CDNet framework, a qualitative comparison was conducted. This analysis specifically evaluates the generalization capabilities of different model variants when transitioning from a synthetic training environment to real-world application scenarios without any domain-specific fine-tuning. The variants under examination incorporate three distinct optical flow backbones: LiteFlowNet, SpyNet, and RAFT, integrated into our dual-branch architecture as Flow-CDNet-L, Flow-CDNet-S, and the proposed Flow-CDNet, respectively.

Fig. 7 illustrates the comparative results on a series of real-world dam bank image pairs, which feature subtle, gradual changes characteristic of slow surface deformation. Each model, having been trained exclusively on the synthetic Flow-Change dataset, was tasked with identifying these changes to assess its cross-domain robustness.

The visual evidence indicates a marked disparity in performance among the backbones. The Flow-CDNet-L variant, which employs LiteFlowNet, struggles significantly with localization. Its output is characterized by diffuse, noisy predictions that fail to form a coherent change mask, indicating that the model's architectural limitations prevent it from generalizing effectively across different data domains without further training.

In contrast, the Flow-CDNet-S model, utilizing SpyNet, demonstrates a moderate improvement. It successfully localizes the general area of change, producing more compact detection masks. However, its performance is compromised by the presence of visual artifacts and a lower degree of shape fidelity when compared to the ground truth. The resulting detections lack the clean boundaries and structural accuracy required for high-fidelity change analysis.

The superior performance of the RAFT-based Flow-CDNet is evident. This configuration consistently generates change masks that are both structurally coherent and spatially precise, aligning closely with the ground truth annotations. The model effectively suppresses noise and preserves the fine-grained boundaries of the deforming regions. This robust generalization from synthetic data to complex, real-world scenes underscores the advanced capabilities of the RAFT architecture in capturing intricate motion cues. The results validate the selection of RAFT as the foundational optical flow branch, as its capacity for accurate and detailed motion estimation is critical to enhancing the change detection accuracy of the unified framework.

4.5.2 Effect of Flow Estimation on Change Detection Accuracy

We examine the influence of the optical flow estimation branch on the performance of the change detection branch. Fig. 8 provides a visual comparison of the output from three different Flow-CDNet configurations on real-world data, highlighting their ability to detect rapid changes. Each model was trained exclusively on a synthetic dataset, thereby testing the generalization of the optical flow backbone's estimated motion to drive the change detection task in a new, un-seen domain.

The results reveal a clear dependency between the quality of the estimated optical flow and the precision of the final change mask. The Flow-CDNet-L variant, which incorporates the LiteFlowNet backbone for flow estimation, demonstrates significant limitations. The change detection branch, relying on the low-quality motion information from this backbone, produces fragmented and noisy masks that fail to accurately delineate the changed regions. This

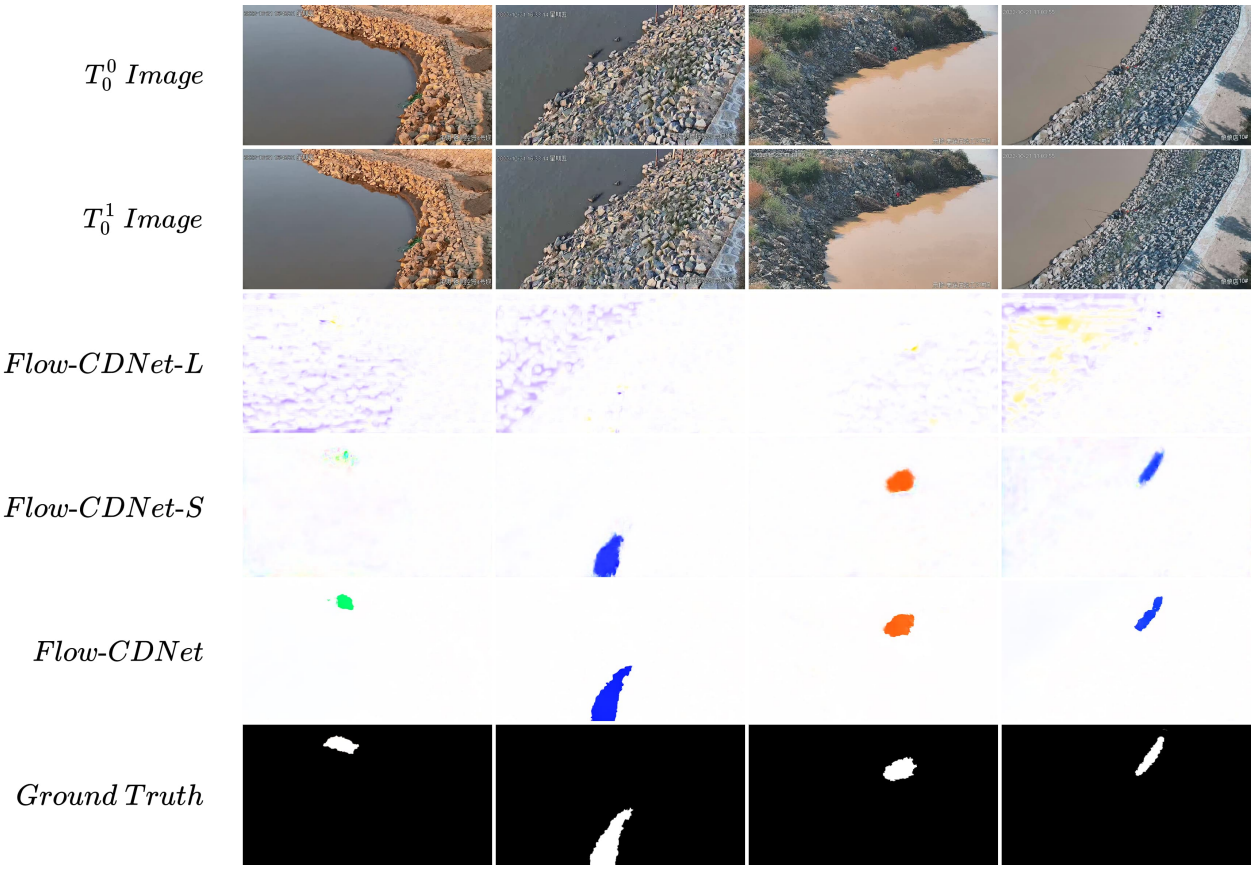

Figure 7: Visual comparison of slow change detection performance using different optical flow backbones on real-world data. From top to bottom, the rows display: the input image at time T_0^0, the input image at time T_0^1, the output from Flow-CDNet-L (LiteFlowNet), Flow-CDNet-S (SpyNet), the proposed Flow-CDNet (RAFT), and the ground truth mask.

indicates that the LiteFlowNet-based flow estimation does not generalize well from synthetic training data, subsequently hampering the downstream change detection task.

In contrast, the Flow-CDNet-S model, which uses SpyNet for optical flow estimation, provides a moderately improved input to the change detection branch. This results in more consolidated detection masks and better localization of changes. However, as depicted in the third and fourth columns, the flow information is still inaccurate, leading to fragmented change masks that miss significant portions of the change region. This illustrates that while the SpyNet-based flow is better than LiteFlowNet's, it is still insufficient to fully support a robust change detection performance.

The proposed Flow-CDNet, leveraging the RAFT architecture for optical flow estimation, consistently provides the most reliable input to the change detection branch. The resulting change masks are structurally coherent, spatially precise, and align closely with the ground truth annotations. The RAFT backbone's superior ability to capture detailed and accurate motion cues from one domain and apply them to another directly translates to enhanced change detection accuracy. The robust generalization of the RAFT-based flow estimation is instrumental in suppressing noise and preserving the fine-grained boundaries of the changing regions, confirming its critical role in boosting the overall performance of the unified framework.

5 Conclusion

A novel change detection framework called Flow-CDNet is proposed, which can simultaneously detect slow and fast changes. To train and evaluate this new framework, we build a merged change detection dataset namely Flow-Change, and design a loss function combining binary tversky loss and L2 norm loss, along with a new evaluation metric called FEPE. Evaluations on real-world data indicate that this framework demonstrates robust detection capabilities in

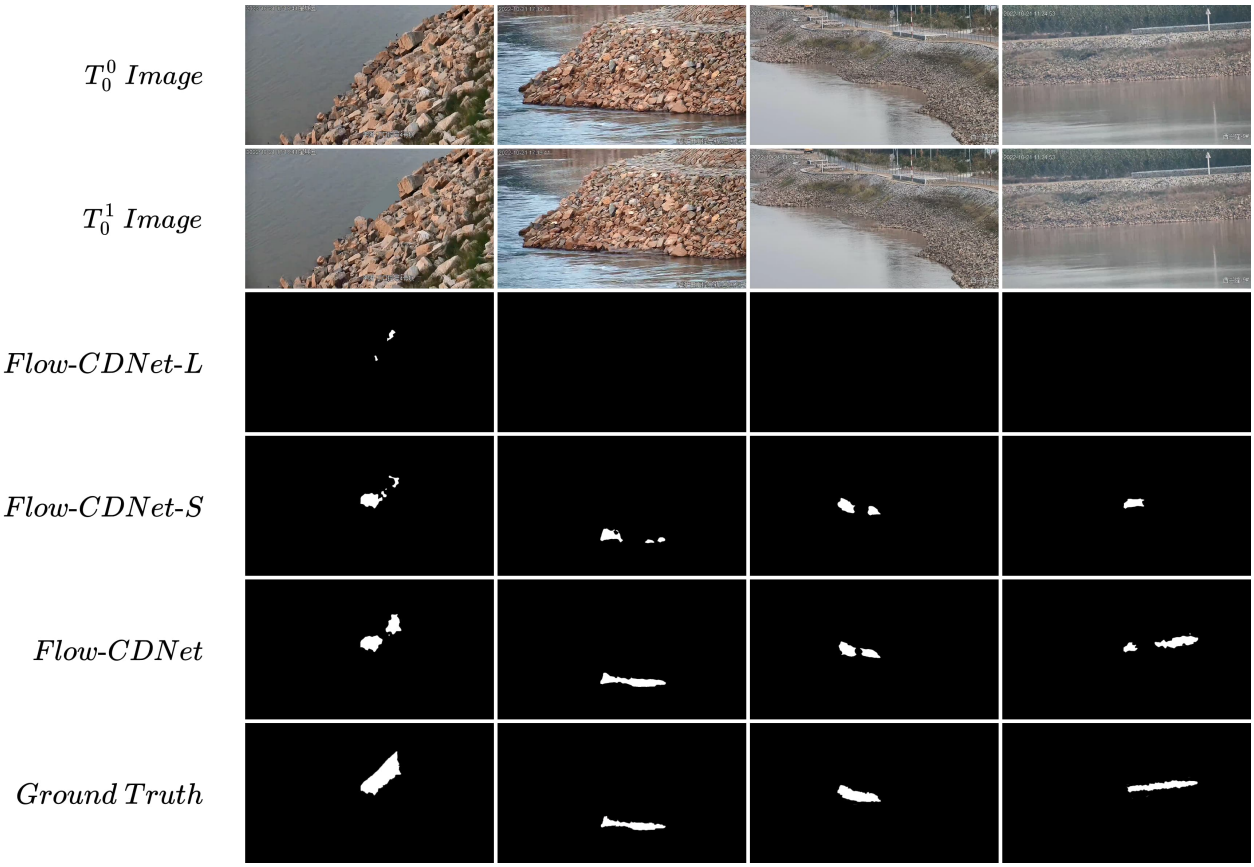

Figure 8: Visual comparison of fast change detection performance with different optical flow backbones on real-world data. From top to bottom, the rows display: the input image at time T_0^0, the input image at time T_0^1, the output from Flow-CDNet-L (LiteFlowNet), Flow-CDNet-S (SpyNet), the proposed Flow-CDNet (RAFT), and the ground truth mask.

authentic scenarios characterized by both gradual and swift changes. Furthermore, through ablation experiments, we verify that the two network branches mutually enhance Flow-CDNet's detection performance. In the future, we'll focus on more challenging scenarios such as defocus, blurriness, or significant lighting variations (day and night).

References

[1] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4161–4170, 2017.

[2] Michal Neoral, Jan Šochman, and Jiří Matas. Continual occlusion and optical flow estimation. In *Asian conference on computer vision*, pages 159–174. Springer, 2018.

[3] Shengyu Zhao, Yilun Sheng, Yue Dong, Eric I Chang, Yan Xu, et al. Maskflownet: Asymmetric feature matching with learnable occlusion mask. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6278–6287, 2020.

[4] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8981–8989, 2018.

[5] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.

[6] Guangyang Wu, Xiaohong Liu, Kunming Luo, Xi Liu, Qingqing Zheng, Shuaicheng Liu, Xinyang Jiang, Guangtao Zhai, and Wenyi Wang. Accflow: Backward accumulation for long-range optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12119–12128, 2023.

[7] Xiaoyu Shi, Zhaoyang Huang, Weikang Bian, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Videoflow: Exploiting temporal cues for multi-frame optical flow estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12469–12480, 2023.

[8] 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.

[9] Jie Chen, Ziyang Yuan, Jian Peng, Li Chen, Haozhe Huang, Jiawei Zhu, Yu Liu, and Haifeng Li. Dasnet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:1194–1206, 2020.

[10] Yi Liu, Chao Pang, Zongqian Zhan, Xiaomeng Zhang, and Xue Yang. Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model. *IEEE Geoscience and Remote Sensing Letters*, 18(5):811–815, 2020.

[11] Jerome Revaud, Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. Epicflow: Edge-preserving interpolation of correspondences for optical flow. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1164–1172, 2015.

[12] Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, and Cordelia Schmid. Deepflow: Large displacement optical flow with deep matching. In *Proceedings of the IEEE international conference on computer vision*, pages 1385–1392, 2013.

[13] Junhwa Hur and Stefan Roth. Mirrorflow: Exploiting symmetries in joint optical flow and occlusion estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 312–321, 2017.

[14] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[15] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017.

[16] Tak-Wai Hui and Chen Change Loy. Liteflownet3: Resolving correspondence ambiguity for more accurate optical flow estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 169–184. Springer, 2020.

[17] Shiyu Zhao, Long Zhao, Zhixing Zhang, Enyu Zhou, and Dimitris Metaxas. Global matching with overlapping attention for optical flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17592–17601, 2022.

[18] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8121–8130, 2022.

[19] Shihao Jiang, Yao Lu, Hongdong Li, and Richard Hartley. Learning optical flow from a few matches. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16592–16600, 2021.

[20] Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer: A transformer architecture for optical flow. In *European conference on computer vision*, pages 668–685. Springer, 2022.

[21] Shili Zhou, Ruian He, Weimin Tan, and Bo Yan. Samflow: Eliminating any fragmentation in optical flow with segment anything model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7695–7703, 2024.

[22] Hyunyoung Jung, Zhuo Hui, Lei Luo, Haitao Yang, Feng Liu, Sungjoo Yoo, Rakesh Ranjan, and Denis Demandolx. Anyflow: Arbitrary scale optical flow with implicit neural representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5455–5465, 2023.

[23] Jisoo Jeong, Hong Cai, Risheek Garrepalli, and Fatih Porikli. Distractflow: Improving optical flow estimation via realistic distractions and pseudo-labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13691–13700, 2023.

[24] Henrique Morimitsu, Xiaobin Zhu, Roberto M Cesar, Xiangyang Ji, and Xu-Cheng Yin. Rapidflow: Recurrent adaptable pyramids with iterative decoding for efficient optical flow estimation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2946–2952. IEEE, 2024.

[25] Wonyong Seo, Woonsung Park, and Munchurl Kim. Lightweight optical flow estimation using 1d matching. *IEEE Access*, 2024.

[26] Qiaole Dong, Chenjie Cao, and Yanwei Fu. Rethinking optical flow from geometric matching consistent perspective. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 1337–1347, 2023.

[27] Xiuchao Sui, Shaohua Li, Xue Geng, Yan Wu, Xinxing Xu, Yong Liu, Rick Goh, and Hongyuan Zhu. Craft: Cross-attentional flow transformer for robust optical flow. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 17602–17611, 2022.

[28] Shunpan Liang, Xirui Zhang, and Yulei Hou. I-raft: Optical flow estimation model based on multi-scale initialization strategy. In *International Conference on Neural Information Processing*, pages 16–29. Springer, 2023.

[29] Ao Luo, Fan Yang, Xin Li, and Shuaicheng Liu. Learning optical flow with kernel patch attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8906–8915, 2022.

[30] Ao Luo, Xin Li, Fan Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Flowdiffuser: Advancing optical flow estimation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19167–19176, 2024.

[31] Afsana Ahsan Jeny, Md Baharul Islam, and Tarkan Aydin. Deeppynet: A deep feature pyramid network for optical flow estimation. In *2021 36th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6. IEEE, 2021.

[32] Ahmed Hammad Alhawwary, Janne Mustaniemi, and Janne Heikkila. Patchflow: A two-stage patch-based approach for lightweight optical flow estimation. In *Proceedings of the Asian Conference on Computer Vision*, pages 3740–3756, 2022.

[33] Shaojie Bai, Zhengyang Geng, Yash Savani, and J Zico Kolter. Deep equilibrium optical flow estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 620–630, 2022.

[34] Stefano Savian, Pietro Morerio, Alessio Del Bue, Andrea A Janes, and Tammam Tillo. Towards equivariant optical flow estimation with deep learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5088–5097, 2023.

[35] Xiuwei Zhang, Yizhe Yang, Lingyan Ran, Liang Chen, Kangwei Wang, Lei Yu, Peng Wang, and Yanning Zhang. Remote sensing image semantic change detection boosted by semi-supervised contrastive learning of semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[36] Xiuwei Zhang, Yuanzeng Yue, Wenxiang Gao, Shuai Yun, Qian Su, Hanlin Yin, and Yanning Zhang. Difunet++: A satellite images change detection network based on unet++ and differential pyramid. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021.

[37] Xiuwei Zhang, Mu Tian, Yinghui Xing, Yuanzeng Yue, Yanping Li, Hanlin Yin, Runliang Xia, Jin Jin, and Yanning Zhang. Adhr-cdnet: Attentive differential high-resolution change detection network for remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022.

[38] Lei Ma, Yu Liu, Xueliang Zhang, Yuanxin Ye, Gaofei Yin, and Brian Alan Johnson. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS journal of photogrammetry and remote sensing*, 152:166–177, 2019.

[39] Changdong Yu, Xiaojun Bi, and Yiwei Fan. Deep learning for fluid velocity field estimation: A review. *Ocean Engineering*, 271:113693, 2023.

[40] 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.

[41] Rodrigo Caye Daudt, Bertr Le Saux, and Alexandre Boulch. Fully convolutional siamese networks for change detection. In *2018 25th IEEE international conference on image processing (ICIP)*, pages 4063–4067. IEEE, 2018.

[42] Ozan Tezcan, Prakash Ishwar, and Janusz Konrad. Bsuv-net: A fully-convolutional neural network for background subtraction of unseen videos. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2774–2783, 2020.

[43] XF Ou, Peng-cheng YAN, Han-pu WANG, et al. Research of moving object detection based on deep frame difference convolution neural network. *Acta Electron. Sin*, 48(12):2384–2393, 2020.

[44] Qin Cheng, Ziliang Ren, Jianming Liu, and Jun Cheng. Multiple time scale motion images for action recognition. In *2020 IEEE International Conference on E-health Networking, Application & Services (HEALTHCOM)*, pages 1–5. IEEE, 2021.

[45] Dalong Zheng, Zebin Wu, Jia Liu, Yang Xu, Chih-Cheng Hung, and Zhihui Wei. Explicit change-relation learning for change detection in vhr remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 2024.

[46] Yinghui Xing, Qi Zhang, Lingyan Ran, Xiuwei Zhang, Hanlin Yin, and Yanning Zhang. Progressive modality-alignment for unsupervised heterogeneous change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–12, 2023.

[47] Yuqun Yang, Xu Tang, Fang Liu, Jingjing Ma, and Licheng Jiao. Remote sensing image change detection based on deep dictionary learning. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 1416–1419. IEEE, 2022.

[48] Mengxuan Zhang, Zhao Liu, Jie Feng, Licheng Jiao, and Long Liu. Deep siamese network with contextual transformer for remote sensing images change detection. In *International Conference on Intelligence Science*, pages 193–200. Springer, 2022.

[49] Daobo Sun, Haohao Yu, Tao Dong, Xiangxu Meng, and Bin Tang. A hybrid method for remote sensing change detection. In *2023 2nd International Conference on Artificial Intelligence, Human-Computer Interaction and Robotics (AIHCIR)*, pages 215–219. IEEE, 2023.

[50] Chang Wang, Shijing Han, Wen Zhang, and Shufeng Miao. Building change detection using deep learning for remote sensing images. *Journal of Information Processing Systems*, 18(4):587–598, 2022.

[51] Josephina Paul. Change detection by deep learning models. In *2022 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*, pages 323–326. IEEE, 2022.

[52] Wei Nie, Peng Gou, Yang Liu, Bhaskar Shrestha, Tianyu Zhou, Nuo Xu, Peng Wang, and QiQi Du. Semi supervised change detection method of remote sensing image. In *2022 IEEE 6th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pages 1013–1019. IEEE, 2022.

[53] Duowang Zhu, Xiaohu Huang, Haiyan Huang, Zhenfeng Shao, and Qimin Cheng. Changevit: Unleashing plain vision transformers for change detection. *arXiv preprint arXiv:2406.12847*, 2024.

[54] Sijun Dong, Libo Wang, Bo Du, and Xiaoliang Meng. Changeclip: Remote sensing change detection with multimodal vision-language representation learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 208:53–69, 2024.

[55] Kanghua Yin, Fang Liu, Jia Liu, and Liang Xiao. Vision-language joint learning for box-supervised change detection in remote sensing. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 10254–10258. IEEE, 2024.

[56] Jiangwei Chen, Sijun Dong, and Xiaoliang Meng. Sganet: A siamese geometry-aware network for remote sensing change detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.

[57] Jiangtao Meng, Xinying Xu, Pengyue Li, and Zhe Zhang. Changead: Enhanced remote sensing change detection via bi-temporal alignment and differential feature integration. In *2024 China Automation Congress (CAC)*, pages 5391–5396. IEEE, 2024.

[58] Lhuqita Fazry, Mgs M Luthfi Ramadhan, and Wisnu Jatmiko. Improving remote sensing change detection via locality induction on feed-forward vision transformer. *Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)*, 16(2), 2023.

[59] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In *International workshop on machine learning in medical imaging*, pages 379–387. Springer, 2017.

[60] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111:98–136, 2015.