FuTH-Net: Fusing Temporal Relations and Holistic Features for Aerial Video Classification

Pu Jin, *Member, IEEE*, Lichao Mou, Yuansheng Hua, *Graduate Student Member, IEEE*, Gui-Song Xia, *Senior Member, IEEE*, and Xiao Xiang Zhu, *Fellow, IEEE*

***Abstract*— Unmanned aerial vehicles (UAVs) are now widely applied to data acquisition due to its low cost and fast mobility. With the increasing volume of aerial videos, the demand for automatically parsing these videos is surging. To achieve this, current research mainly focuses on extracting a holistic feature with convolutions along both spatial and temporal dimensions. However, these methods are limited by small temporal receptive fields and cannot adequately capture long-term temporal dependencies that are important for describing complicated dynamics. In this article, we propose a novel deep neural network, termed Fusing Temporal relations and Holistic features for aerial video classification (FuTH-Net), to model not only holistic features but also temporal relations for aerial video classification. Furthermore, the holistic features are refined by the multiscale temporal relations in a novel fusion module for yielding more discriminative video representations. More specially, FuTH-Net employs a two-pathway architecture: 1) a holistic representation pathway to learn a general feature of both frame appearances and short-term temporal variations and 2) a temporal relation pathway to capture multiscale temporal relations across arbitrary frames, providing long-term temporal dependencies. Afterward, a novel fusion module is proposed to spatiotemporally integrate the two features learned from the two pathways. Our model is evaluated on two aerial video classification datasets, ERA and Drone-Action, and achieves the state-of-the-art results. This demonstrates its effectiveness and good generalization capacity across different recognition tasks (event classification and human action recognition). To facilitate further research, we release the code at https://gitlab.lrz.de/ai4eo/reasoning/futh-net.**

***Index Terms*— Aerial video classification, convolutional neural networks (CNNs), holistic features, temporal relations, two-pathway, unmanned aerial vehicle (UAV).**

Manuscript received August 23, 2021; revised November 5, 2021; accepted December 23, 2021. Date of publication February 10, 2022; date of current version March 24, 2022. This work was supported in part by the European Research Council (ERC) through the European Union's Horizon 2020 Research and Innovation Programme [10^{16} Bytes from Social Media to Earth Observation Satellites (So2Sat)] under Grant ERC-2016-StG-714087; in part by the Helmholtz Association through the Framework of Helmholtz AI under Grant ZT-I-PF-5-01; in part by the Local Unit "Munich Unit @Aeronautics, Space and Transport (MASTr)" and Helmholtz Excellent Professorship "Data Science in Earth Observation—Big Data Fusion for Urban Research" under Grant W2-W3-100; in part by the German Federal Ministry of Education and Research (BMBF) in the Framework of the international future AI lab "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond under Grant 01DD20001; and in part by the German Federal Ministry of Economics and Technology in the Framework of the "National Center of Excellence ML4Earth" under Grant 50EE2201C. *(Corresponding authors: Lichao Mou; Xiao Xiang Zhu.)*

Pu Jin is with the State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan 430072, China, and also with the Department of Aerospace and Geodesy, Technical University of Munich (TUM), 80333 Munich, Germany (e-mail: pu.jin@tum.de).

Lichao Mou, Yuansheng Hua, and Xiao Xiang Zhu are with the Remote Sensing Technology Institute, German Aerospace Center, 82234 Weßling, Germany, and also with the Data Science in Earth Observation (Signal Processing in Earth Observation), Technical University of Munich, 80333 Munich, Germany (e-mail: lichao.mou@dlr.de; yuansheng.hua@dlr.de; xiaoxiang.zhu@dlr.de).

Gui-Song Xia is with the State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS) and the School of Computer Science, Wuhan University, Wuhan 430072, China (e-mail: guisong.xia@whu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2022.3150917

I. Introduction

BY THE virtue of low-cost, real-time, and high-resolution data acquisition capacity, unmanned aerial vehicles (UAVs) can be exploited for a wide range of applications [1]–[17] in the field of remote sensing, such as object tracking and surveillance [5]–[10], traffic flow monitoring [11]–[14], and precision agriculture [15]–[17]. With the proliferation of UAVs worldwide, the number of produced aerial videos is significantly increasing. Hence, there is an escalating demand for automatically parsing aerial videos because it is unrealistic for humans to screen such big data and understand their contents. Therefore, aerial video classification becomes an important task in aerial video interpretation [18].

Feature learning and representation from videos is crucial for this task. Convolutional neural networks (CNNs) have demonstrated the superb capability of learning effective visual representations from images. For instance, ResNet [19] has achieved an impressive performance on the ImageNet dataset, which is even better than the reported human-level performance [20]. Compared to a sequence of remote sensing images in which the temporal information is limited due to relatively long satellite revisit periods, an overhead video is able to deliver more fine-grained temporal dynamics that are essential for describing complex events. Therefore, moving from image recognition to video classification, much effort has been made to learn spatiotemporal feature representations.

On the one hand, several methods [21]–[38] aim at learning a global spatiotemporal feature representation that can holistically represent a video. A straightforward idea is to extract spatiotemporal features on each video frame individually by making use of 2-D convolutions and then pool stacked feature maps across the temporal domain [21]. However, this could lead to the ignorance of temporal relations among various frames. To address this, Donahue *et al.* [22] and Ng *et al.* [23] employed recurrent neural networks (RNNs), such as long short-term memory (LSTM) [39] to model temporal relations by integrating features over time. However,

the effectiveness of such methods usually depends heavily on the learning effect of long-term memorization. Furthermore, 3-D CNNs are fairly natural models for video representation learning and able to learn global spatiotemporal features by performing 3-D convolutions in both spatial and temporal dimensions. Some 3-D CNN architectures [31]–[38] have been investigated and shown impressive performance. For instance, Tran *et al.* [31] proposed a 3-D CNN model with $3 \times 3 \times 3$ convolution filters for learning a video representation on a large-scale video dataset. Nonetheless, massive computational consumption and memory demand hinder efforts to train a very deep 3-D CNN and limit the performance of 3-D CNN architectures. To address this problem, inflated 3-D convolution filters [35] and decomposed 3-D convolution filters [36], [37] utilize a more economic method to implement 3-D convolutions and boost the performance of 3-D CNNs. However, the aforementioned methods with either 2-D or 3-D convolutions have limited temporal receptive fields and therefore cannot adequately capture variable temporal dependencies. On the other hand, a few recent works attempt to explicitly model temporal relationships and demonstrate promising results in several tasks, to name a few, temporal relational reasoning [40]–[44], object detection and tracking [6]–[9], event recognition [45]–[47], video segmentation [48]–[50], dynamic texture recognition [51], and spatiotemporal learning [52], [53].

A video delivers not only spatial information but also temporal dynamics. Hence, some studies are dedicated to capture spatial (appearance) and temporal (motion) representations separately by a two-stream architecture. In these two-stream models, fusing the features from two pathways is an important procedure for recognition. For example, Simonyan and Zisserman [24] directly fused the softmax scores using either averaging operation or a simple linear SVM. Karpathy *et al.* [21] utilized a fully connected layer to merge the two streams of the late fusion model. However, its performance is surpassed by a purely spatial network. In addition, Feichtenhofer *et al.* [54] introduced residual connections between appearance and temporal streams to enable motion interactions. For stream fusion, they average the prediction scores of the classification layers from two streams. Feichtenhofer *et al.* [30] investigated several fusion methods such as max, concatenation, convolution, and observe that 3-D convolutional fusion outperforms averaging the softmax output. The main limitation of the two-stream architecture is that it is not capable to spatiotemporally match spatial and temporal information. Therefore, a fusion method is needed to spatiotemporally register the features from two pathways. However, the abovementioned fusion methods leverage a single operation (e.g., averaging) that is not able to effectively enable spatiotemporal interactions between them.

The motion in aerial videos usually has different durations and shows high variability. For example, in the ERA dataset [18], mudslide shows a simple and repeated motion over a long duration, which could be described by a few video frames; car racing depicts a complicated, dynamic process and is composed of a variety of consecutive motions, including chasing, approaching, away, and colliding, over a short duration. Temporal relations across multiple frames are an important cue to represent the complex motion. The aforementioned approaches based on spatiotemporal convolutions (e.g., $3 \times 3 \times 3$ convolutions) simply add a temporal dimension to 2-D convolution filters to implicitly learn temporal dependencies, and they are not adaptable to capture various, complicated temporal dynamics over a long duration due to their limited temporal receptive fields. To address this issue, we propose to explicitly learn temporal relations across arbitrary frames to effectively model long-term temporal dependencies. Furthermore, we introduce multiscale temporal relations into holistic features to design a two-pathway architecture for aerial video classification. Besides, for spatiotemporal registering temporal relations and holistic features, we propose a novel fusion module in which holistic features are spatiotemporally modulated with temporal relations.

In this article, we present a two-pathway network, termed Fusing Temporal relations and Holistic features for aerial video classification (FuTH-Net). One pathway is devised to capture a holistic feature describing appearances and short-term temporal variations. The other pathway is responsible for excavating temporal relations across arbitrary frames at multiple timescales, providing long-term temporal dependencies. Last but not least, for spatiotemporally fusing two features from two pathways, we further present a novel fusion module in which the multiscale temporal relations are leveraged to refine the temporal features in the holistic representation. More specifically, we learn the holistic feature by treating a video as an entirety and using inflated 3-D convolution operators [35]. Meanwhile, we sample frame-level feature vectors at different sampling rates to learn multiscale temporal relations with a sequence of multilayer perceptrons (MLPs) [55]. As to the fusion of these two features, we employ a fusion module in which the temporal relations are modulated with the holistic representation by a normalization-like process [56], [57]. The resulting feature representation is then fed into the following layers for the purpose of video classification. Contributions of this article are threefold.

1) We propose a novel network, namely FuTH-Net, for the task of aerial video classification. This network exploits a two-pathway architecture, one for learning a video presentation holistically and the other for fully excavating useful temporal relations at multiple timescales among video frames.
2) A novel fusion module exploits a normalization-like pipeline in which the two features learned from two pathways are spatiotemporally registered by modulating the holistic features according to temporal relations. In this module, the temporal information in holistic features is refined by multiscale temporal relations. A more discriminative fused feature is obtained for distinguishing different video events.
3) We evaluate the effectiveness of the proposed network through extensive experiments, and experimental results show that our method achieves the state-of-the-art performance.

The remaining sections of this article are organized as follows. Section II details the architecture of FuTH-Net, and

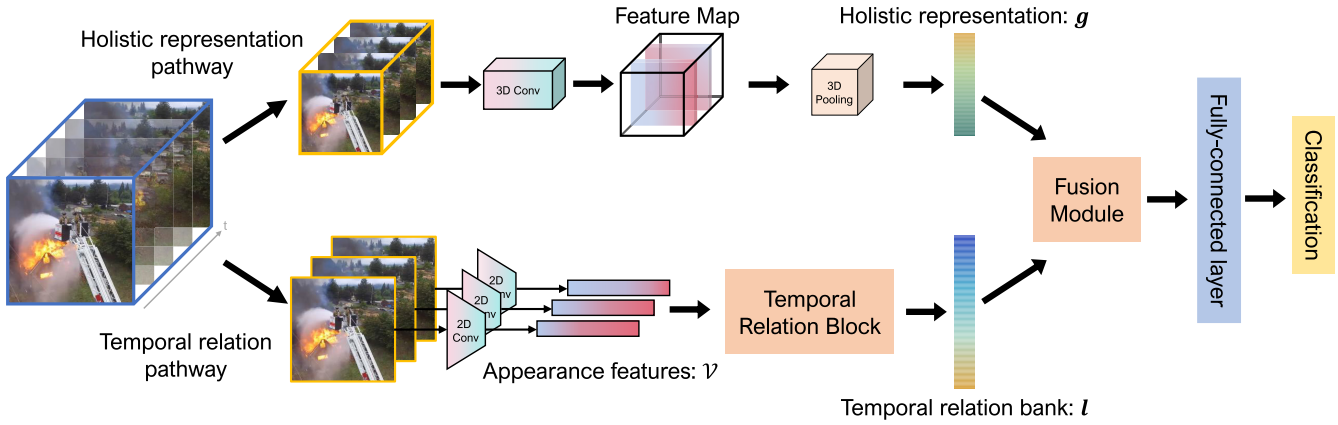


Fig. 1. Overview of FuTH-Net. The upper pathway, namely the holistic representation pathway, aims at capturing a holistic feature $\boldsymbol{g}$ by 3-D convolutions. The lower pathway, namely temporal relation pathway, aims to learn a multiscale temporal relation bank $\boldsymbol{l}$ by a temporal relation block. A followed fusion module combines the outputs of two pathways to generate a robust fused feature $\boldsymbol{z}$, which is finally fed into a fully connected layer for aerial video classification.

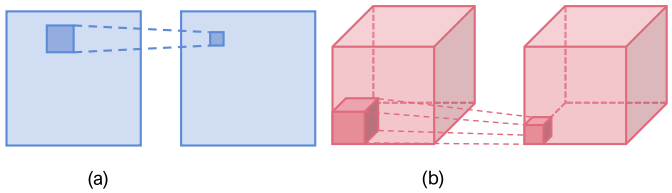


Fig. 2. (a) 2-D convolution versus (b) 3-D convolution. Compared to 2-D convolution, 3-D convolution slides in both temporal and spatial dimensions and results in an output volume, which thereby captures both spatial and temporal information, i.e., the holistic representation.

Section III shows and discusses experimental results. Finally, the conclusion is drawn in Section IV.

II. Network Architecture

In this section, we detail our proposed network architecture, FuTH-Net, for aerial video classification. First, we introduce an overview of the proposed network in Section II-A. Furthermore, we give more detailed descriptions for two modules, temporal relation block and fusion module, in Sections II-B and II-C, respectively. Finally, the implementation of our network is introduced in Section II-D.

A. FuTH-Net

The motivation of our network is to simultaneously model the holistic feature and temporal relations of a video with a two-pathway architecture. The resulting two feature representations are integrated by a fusion module. The overview of the architecture is shown in Fig. 1.

Holistic representation pathway treats a video as an entity and aims at learning a holistic feature by 3-D convolutions. The 3-D convolution is achieved by endowing 2-D convolution with an additional dimension (e.g., the temporal dimension of aerial videos), which is shown in Fig. 2. Compared to 2-D convolution, 3-D convolution is able to capture both spatial information and temporal information, so-called holistic representation in our case. It is of importance for video classification under some circumstances where events with simple temporal dynamics are strongly associated with certain objects or scenes. As to the implementation of 3-D convolutions, many efforts, e.g., 3-D convolutional kernel [31], inflated 3-D convolution [35], and pseudo-3-D convolution [36], have been made to symmetrically extract both spatial information and short-term temporal information. In this work, we choose a typical 2-D CNN architecture and transform all 2-D operations to 3-D operations by a specific 3-D implementation method [35]. Then, we employ the transformed 3-D CNN with a bunch of 3-D convolution and pooling operations on a video volume to capture a holistic representation $\boldsymbol{g}$.

Temporal relation pathway views a video as a sequence of frames and aims to capture temporal relations across multiple frames by a temporal relation block. Temporal relation information is vital for video classification, as it is capable of capturing high-level interactions among entities (subjects, objects, scenes, and so on) over a long temporal series, which are significant for recognizing events with complex temporal dynamics. To take advantage of this cue, we apply a 2-D CNN to video frames to extract appearance features. Then, these features are fed into the temporal relation block to learn a multiscale temporal relation bank $\boldsymbol{l}$ across arbitrary frames.

Fusion module combines the outputs of the two pathways to build a more discriminative representation. More specifically, it leverages a normalization-like pipeline in which the temporal relations are transformed to two modulation parameters by two affine transformations, and the produced parameters $\mathcal{F}_1(\boldsymbol{l})$ $\mathcal{F}_2(\boldsymbol{l})$ multiplied and added with the holistic feature $\boldsymbol{g}$ to yield the normalized activation elementwisely. Finally, the fused feature $\boldsymbol{z}$ is obtained by concatenating the normalized activation with an additional holistic feature $\boldsymbol{g}$.

In what follows, we detail the temporal relation block and fusion module.

B. Temporal Relation Block

The purpose of temporal relational reasoning lies in linking meaningful transformations among entities over time. The work [58] is intended to construct a fully connected

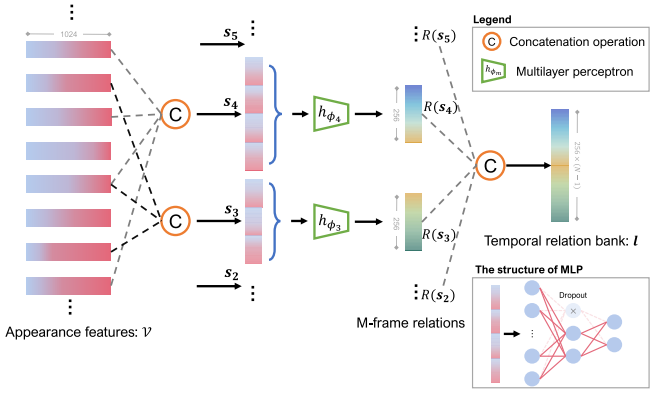


Fig. 3. Temporal relation block. Appearance feature vectors are randomly selected from the feature set $\mathcal{V}$. Afterward, the selected vectors are concatenated and then fed into an MLP to learn a corresponding m-frame relation. Finally, all m-frame relations are concatenated to produce a multiscale temporal relation bank $\boldsymbol{l}$.

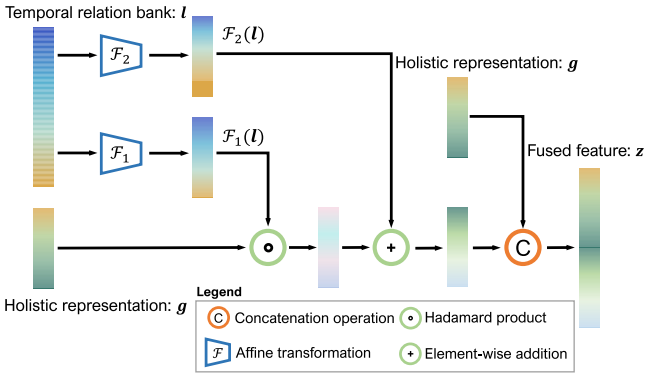


Fig. 4. Fusion module. Two affine transformations are applied on the temporal relation bank $\boldsymbol{l}$ to produce two vectors, $\mathcal{F}_1(\boldsymbol{l})$ and $\mathcal{F}_2(\boldsymbol{l})$. Afterward, the Hadamard product and addition operation are applied on them with $\boldsymbol{g}$. Finally, the output vector is concatenated with $\boldsymbol{g}$ to yield the fused feature $\boldsymbol{z}$.

graph among entities in video frames and calculate pairwise energy functions among node pairs in the graph to model temporal relations. Inspired by this work, we aim at capturing temporal relations among arbitrary frames. Instead of utilizing a fully connected graph among video frames which inevitably increases computation and redundancy, we make use of a sampling strategy to sample multiple snippets and learn relational representations using a group of MLPs. Note that each sampled snippet contains a variable number of frames for the purpose of learning multiscale relational representations.

Formally, suppose that we have extracted an appearance feature set $\mathcal{V} = \{\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_N\}$ of video frames by a 2-D CNN, where $\boldsymbol{f}_i$ denotes the 256-D feature vector of the ith video frame and N is the number of frames. We randomly sample m vectors from $\mathcal{V}$ and concatenate them to $\boldsymbol{s}_m$, where m is the total number of sampled frames and the length of vector $\boldsymbol{s}_m$ is $m \times 256$. Notably, before concatenation, we rearrange sampled vectors according to the original temporal order. The corresponding m-frame relation function is defined as follows:

$$\mathrm{R}(\boldsymbol{s}_m) = h_{\phi_m}(\boldsymbol{s}_m) \tag{1}$$

where the input is the concatenated vector $\boldsymbol{s}_m = [\boldsymbol{f}_i, \boldsymbol{f}_j, \ldots, \boldsymbol{f}_p]$, $i, j, p \in [1, N]$, $m \in [2, N]$, and $[\cdot, \cdot]$ denotes concatenation. h_{ϕ_m} is a two-layer MLP with parameters ϕ_m and learns the transformations among m feature vectors. The parameters of h_{ϕ_m} are learned separately with respect to each $\boldsymbol{s}_m$. With variant values m, temporal relations at multiple timescales can be yielded and further concatenated to build a multiscale temporal relation bank $\boldsymbol{l} = [\mathrm{R}(\boldsymbol{s}_2), \mathrm{R}(\boldsymbol{s}_3), \ldots, \mathrm{R}(\boldsymbol{s}_N)]$.

The temporal relation block is a basic computational unit with an input feature set $\mathcal{V}$ and an output temporal relation bank $\boldsymbol{l}$ and can be easily plugged into any classification CNN models. Fig. 3 shows the structure of our temporal relation block.

C. Fusion Module

Outputs from the holistic representation pathway and temporal relation pathway are integrated by a fusion module that encodes spatiotemporal correspondences between holistic features and temporal relations. Spatiotemporally registering the two features is vital for encoding spatiotemporal correspondences. Motivated by conditional normalization [56], [57], we present a novel fusion module where the two features are spatiotemporally registered by modulating the holistic features according to temporal relations. The multiscale temporal information is leveraged to refine the temporal representations in holistic features. Specifically, the module utilizes a normalization-like pipeline in which the temporal relations are transformed to two modulation parameters by two affine transformations, and the produced parameters $\mathcal{F}_1(l)$ and $\mathcal{F}_2(l)$ are multiplied and added with $\boldsymbol{g}$ to yield the normalized activation elementwisely. Finally, the fused feature $\boldsymbol{z}$ is obtained by concatenating the normalized activation with an additional holistic feature $\boldsymbol{g}$. The fusion equation is given as follows:

$$\boldsymbol{z} = \left[\mathcal{F}_1(\boldsymbol{l}) \odot \boldsymbol{g} + \mathcal{F}_2(\boldsymbol{l}), \boldsymbol{g}\right] \tag{2}$$

where $\mathcal{F}_1$ and $\mathcal{F}_2$ are affine transformations and aim to produce the modulation parameters, $\odot$ denotes a Hadamard production, and $[\cdot, \cdot]$ denotes concatenation. The overall structure of fusion module is shown in Fig. 4. We concatenate an additional holistic feature $\boldsymbol{g}$ with the modulated feature to yield the final fused feature $\boldsymbol{z}$. This is for enriching the spatial information that is important for distinguishing events with simple dynamics. For validating its effectiveness, we further compare it with several existing fusion methods in ablation study (see Section III-B).

D. Implementation Details

In this section, we describe the implementation of our FuTH-Net.

1) Holistic Representation Pathway: We convert a typical image classification architecture, Inception-v1 [59], into a 3-D architecture by inflating all convolutions and pooling filters. The 3-D convolutions are created by endowing 2-D ones with an additional temporal dimension. Furthermore, we would like to bootstrap the 2-D network weights pretrained on ImageNet into the 3-D model. To achieve this, the 3-D model could be implicitly pretrained on ImageNet by

converting images into fixed videos. We replicate weights of 2-D convolutions N times along the temporal dimension and then divide them by N to produce pretrained parameters for the 3-D model. Moreover, we optimize hyperparameters for convolutions and pooling operations (e.g., stride and pooling size) to effectively capture representative temporal dynamics. In detail, we use $1 \times 3 \times 3$ kernels with $1 \times 2 \times 2$ strides in the first two max-pooling layers for remaining initial temporal information. The final average-pooling layer exploits a $2 \times 7 \times 7$ kernel to produce a 1024-D feature vector that is regarded as the holistic representation $\boldsymbol{g}$.

2) Temporal Relation Pathway: We utilize Inception-v1 with batch normalization pretrained on ImageNet as our feature extraction model to generate a 1024-D feature vector for each frame. Subsequently, a feature bank with the size of $n \times 1024$ for an input video is produced, where n is the number of input video frames. Moreover, ϕ_m is a two-layer MLP with 256 units, and each layer is followed by a batch normalization [56] layer and a ReLU activation function. $(N-1)$ temporal relations are extracted by ϕ_m and then concatenated into the final multiscale temporal relation bank with the dimension of $256 \times (N-1)$. The number of input frames is set to 16 in both two pathways.

3) Fusion Module: Two simple MLPs with dropout operations are exploited to implement the two affine transformations that are employed on $\boldsymbol{l}$ to yield two 1024-D vectors, $\mathcal{F}_1(\boldsymbol{l})$ and $\mathcal{F}_2(\boldsymbol{l})$. The final fused feature is a 2048-D vector.

4) Training Schedule: The network is trained on PyTorch[1] framework and runs on one NVIDIA Tesla P100 GPU[2] with 16-GB on-board memory. We train our model with a stochastic gradient descent (SGD) [60] optimizer using a momentum of 0.9 and a weight decay of 0.0005. Due to the limitation of GPU memory, we utilize a multistage training strategy. Specifically, the whole training procedure is composed of three phases. First, we train the holistic representation pathway for 100 epochs with a batch size of 6 and a learning rate of 0.001. Then, we train the temporal relation pathway with a learning rate of 0.0001 and the same epochs and batch size while keeping weights of the holistic representation pathway fixed. Finally, the fusion module is trained for 120 epochs with weights of two pathways fixed.

III. Experiments

In this section, we first introduce aerial video recognition datasets, competitors, and evaluation metrics in Section III-A. Then, we perform ablation studies to investigate the complementarity between the holistic representation pathway and the temporal relation pathway as well as the effectiveness of our fusion module in Section III-B. Furthermore, we assess the performance of our FuTH-Net on two different aerial video recognition datasets, ERA and Drone-Action, and analyze the experimental results in Sections III-C and III-D, respectively.

A. Experimental Setup

1) Datasets: To evaluate the performance of FuTH-Net, we conduct experiments on two aerial video recognition datasets with standard evaluation protocols. First, we use the ERA dataset [18] that is an event recognition dataset and consists of 2864 aerial event videos collected from YouTube. In this dataset, 25 events are defined, including postearthquake, flood, fire, landslide, mudslide, traffic collision, traffic congestion, harvesting, ploughing, constructing, police chase, conflict, baseball, basketball, boating, cycling, running, soccer, swimming, car racing, party, concert, parade/protest, religious activity, and nonevent (see Fig. 5). Then, the Drone-Action dataset [61] for human action classification in aerial videos is utilized to further assess the performance of models. In this dataset, 240 self-taken aerial videos are collected, and 13 different actions are defined: kicking, walking front/back, running side, jogging side, walking side, hitting stick, running front/back, stabbing, jogging front/back, clapping, hitting bottle, boxing, and waving hands (see Fig. 6). Table I exhibits the details of the two datasets.

TABLE I
DATASET OVERVIEW. WE PROVIDE VARIABLE DETAILS OF THE TWO DATASETS

	ERA Dataset [18]	**Drone-Action Dataset [61]**
Type of Task	general event recognition	human action recognition
Data Source	YouTube	self-collected (actor staged)
# Classes	25	13
Video Size	640×640	1920×1080
Video Duration	5s	5s $\sim$ 21s
# Samples	2864	240

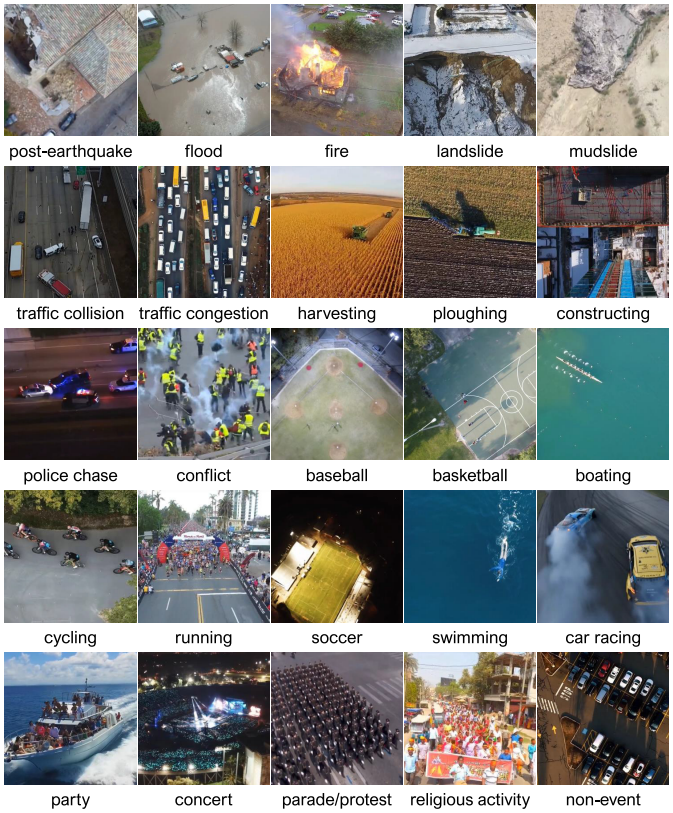


Fig. 5. Overview of the ERA dataset. We show the middle frame of one video in each class.

[1]https://pytorch.org/
[2]https://www.nvidia.com/en-us/data-center/tesla-p100/

TABLE II

COMPARISON WITH STATE-OF-THE-ART METHODS ON THE ERA DATASET. WE SHOW THE PER-CLASS PRECISION AND OVERALL ACCURACY (OA) ON THE TEST SET. THE BEST PRECISION/ACCURACY IS SHOWN IN BOLD

Model	*post-earthquake*	*flood*	*fire*	*landslide*	*mudslide*	*traffic collision*	*traffic congestion*	*harvesting*	*ploughing*	*constructing*	*police chase*	*conflict*	*baseball*	*basketball*	*boating*	*cycling*	*running*	*soccer*	*swimming*	*car racing*	*party*	*concert*	*parade/protest*	*religious activity*	*non-event*	**OA**	κ
C3D†	23.1	24.3	30.9	19.5	32.9	7.00	15.5	27.5	36.1	45.5	50.0	18.2	40.9	37.0	47.5	20.6	12.0	58.3	36.2	16.7	25.8	38.2	37.8	27.5	29.6	30.4	0.21
C3D‡	27.9	56.5	32.7	10.2	23.9	8.30	38.5	42.3	31.1	40.0	51.9	11.1	45.7	48.9	41.9	13.6	9.30	41.9	38.2	18.2	17.4	32.0	28.1	35.8	28.5	31.1	0.23
P3D†-ResNet-199	43.6	65.9	66.7	35.5	48.7	20.0	37.8	77.4	70.8	62.0	**81.6**	22.2	66.7	63.1	55.4	35.6	35.3	76.2	57.4	40.0	54.5	37.5	38.7	47.8	37.4	50.7	0.47
P3D‡-ResNet-199	72.4	76.3	84.8	24.5	38.2	35.6	40.8	56.9	67.4	71.4	57.9	50.0	70.4	78.8	71.7	47.1	60.0	79.5	68.1	40.9	59.1	37.0	49.1	55.9	37.9	53.3	0.51
I3D†-Inception-v1	40.4	63.5	68.9	22.6	46.3	17.6	55.0	61.5	50.0	53.3	73.2	50.0	75.0	69.4	60.7	61.9	53.3	70.8	52.5	50.0	57.1	50.7	40.3	49.0	35.8	51.3	0.48
I3D‡-Inception-v1	60.0	68.1	65.7	29.0	60.4	**51.5**	52.2	67.1	66.7	54.2	64.8	57.9	**85.0**	61.9	**86.4**	75.0	44.4	77.6	64.1	65.2	53.7	50.0	47.8	65.1	43.0	58.5	0.55
TRN†-BNInception	**84.8**	71.4	82.5	51.2	50.0	46.8	**66.7**	68.1	77.4	52.4	70.5	75.0	64.5	67.7	84.0	56.1	55.2	**83.3**	72.9	61.1	62.0	48.9	44.6	62.8	51.1	62.0	0.58
TRN‡-Inception-v3	69.2	87.8	88.9	**65.8**	60.0	44.1	58.3	78.1	**90.7**	70.8	73.3	28.6	83.3	72.7	73.7	60.0	**66.7**	73.6	70.6	63.6	65.1	47.7	42.7	65.1	47.9	64.3	0.60
SlowFast†	70.1	**88.0**	83.3	57.2	67.3	51.4	56.2	68.4	87.6	**82.0**	75.1	**75.5**	40.8	70.3	71.8	61.4	54.7	78.2	72.9	74.3	50.4	70.3	50.6	**65.7**	60.7	64.9	**0.63**
Multigrid†	69.8	71.7	**89.5**	54.7	64.1	47.4	59.4	**78.4**	73.4	69.4	72.4	51.8	63.8	74.7	76.2	**75.2**	52.1	71.1	69.6	67.5	**66.1**	74.4	55.7	62.3	57.4	65.3	0.62
FuTH-Net	72.7	75.5	87.5	57.1	**74.5**	34.0	56.0	76.6	71.2	81.4	76.5	36.0	78.0	**85.4**	80.4	73.6	16.3	64.5	**80.4**	**84.2**	56.0	**89.8**	**65.3**	63.0	**63.9**	**66.8**	**0.63**

[1] C3D† uses pre-trained weights on the Sport1M dataset as initialization; C3D‡ uses pre-trained weights on the UCF101 dataset as initialization.
[2] P3D†-ResNet-199 uses pre-trained weights on the Kinetics dataset as initialization; P3D‡-ResNet-199 uses pre-trained weights on the Kinetics-600 dataset as initialization.
[3] I3D†-Inception-v1 uses pre-trained weights on the Kinetics dataset as initialization; I3D‡-Inception-v1 uses pre-trained weights on Kinetics+ImageNet as initialization.
[4] TRN†-BNInception uses pre-trained weights on the Something-Something V2 dataset as initialization; TRN‡-Inception-v3 uses pre-trained weights on the Moments in Time dataset as initialization.
[5] SlowFast† is trained from random initialization, without using pre-training.
[6] Multigrid† use ImageNet-pre-trained for 3D convolutions inflated from 2D convolutions following common practice.

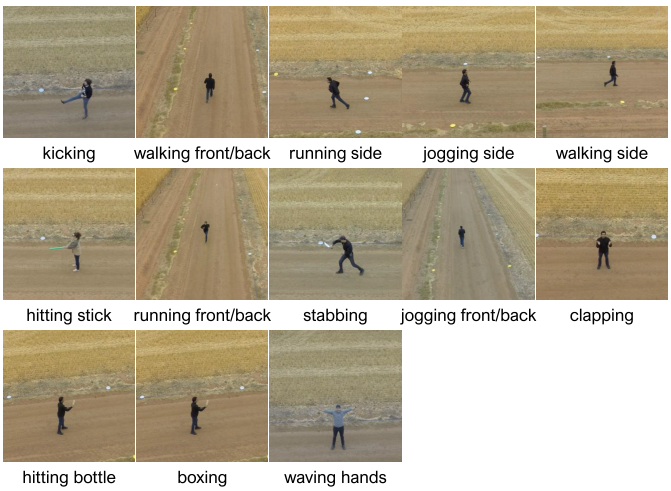


Fig. 6. Overview of the Drone-Action dataset. We show the middle frame of one video in each class.

In the preprocessing phase, we transform video clips of the Drone-Action dataset into the same data structure as the ERA dataset. Since durations of videos in the Drone-Action dataset range from 5 to 21 s, we cut them to 5-s clips. Afterward, each frame is cropped and resized to a size of 640 × 640. For both datasets, we sample 16 frames from each video clip with a fixed sampling rate.

2) Competitors: We compare the proposed network with several state-of-the-art video classification models.

1) *3-D Convolutional Network (C3D) [31]:* It aims to extract spatiotemporal features with 3-D convolutional filters and pooling layers. Compared to conventional 2-D CNNs, 3-D convolutions and pooling operations in C3D can preserve the temporal information of input signals and model motion as well as appearance simultaneously. Moreover, Tran *et al.* [31] demonstrated that the optimal size of 3-D convolutional filters is 3 × 3 × 3. In our experiments, we test two C3D[3] networks with pretrained weights on the Sport1M dataset [21] and the UCF101 dataset [62] (see C3D† and C3D‡ in Table II), respectively.
2) *Pseudo-3-D Residual Network (P3D ResNet) [36]:* It is composed of pseudo-3-D convolutions, where conventional 3-D convolutions are decoupled into 2-D and 1-D convolutions in order to learn spatial and temporal information separately. With such convolutions, the model size of a network can be significantly reduced, and the utilization of pretrained 2-D CNNs is feasible. Besides, inspired by the success of ResNet [19], P3D ResNet employs ResNet-like architectures to learn residuals in both spatial and temporal domains. In our experiments, we test two 199-layer P3D ResNet[4] (P3D-ResNet-199) models with pretrained weights on the Kinetics dataset [63] and the Kinetics-600 dataset [64] (see P3D†-ResNet-199 and P3D‡-ResNet-199 in Table II), respectively.
3) *Inflated 3-D ConvNet (I3D) [35]:* It expands the 2-D convolution and pooling filters to 3-D, which are then initialized with inflated pretrained models. In particular, the weights of 2-D networks pretrained on the ImageNet dataset are replicated along the temporal

[3] https://github.com/tqvinhcs/C3D-tensorflow
[4] https://github.com/zzy123abc/p3d

dimension. With this design, not only 2-D network architectures but also pretrained 2-D models can be efficiently employed to increase the learning efficiency and performance of 3-D networks. To assess the performance of I3D on our dataset, we test two I3D[5] models whose backbones are both Inception-v1 [59] (I3D-Inception-v1) with pretrained weights on the Kinetics dataset [63] and Kinetics+ImageNet (see I3D†-Inception-v1 and I3D‡-Inception-v1 in Table II).

4) *Temporal Relation Network (TRN) [43]:* It is proposed to recognize human actions by reasoning about multiscale temporal relations among video frames. By leveraging the proposed plug-and-play relational reasoning module, TRN can even accurately predict human gestures and human–object interactions through sparsely sampled frames. For our experiments, we test TRNs[6] with 16 multiscale relations and select the inception architecture as the backbone. Notably, we experiment two variants of the inception architecture: BNInception [65] and Inception-v3 [66]. We initialize the former with weights pretrained on the Something-Something V2 dataset [67] (TRN†-BNInception in Table II) and the latter with weights pretrained on the Moments in Time dataset [68] (TRN‡-Inception-v3 in Table II).
5) *SlowFast [46]:* SlowFast network is a two-pathway architecture in which a slow pathway is designed for operating at low frame rate to capture spatial semantic information, and a fast pathway aims at operating at high frame rate to learn motion at fine temporal resolution. To assess the performance of SlowFast on our dataset, we test one SlowFast[7] model (see SlowFast† in Table II) whose backbone is ResNet [19] without pretraining.
6) *Multigrid [47]:* Multigrid training method utilizes variable minibatch shapes with different spatiotemporal resolutions in the training phase. The different shapes are generated by resampling the training data on multiple sampling grids. The novel training method yields a significant out-of-the-box training speedup for different models (I3D, SlowFast). In our experiments, we use this training method test SlowFast network[8] (see Multigrid† in Table II) with ImageNet-pretraining.

3) Evaluation Metrics: We make use of the per-class precision, OA, confusion matrix, and kappa coefficient as evaluation metrics for comparing the performance of different models. Specifically, the preclass precision is calculated with the following equation:

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}. \tag{3}$$

The OA is computed by dividing the number of correctly classified test samples by that of all test samples. Moreover, the confusion matrix is visualized to illustrate the classification performance of variant models. Each element of the matrix denotes the number of instances that belong to the ground-truth class (X-axis) but are classified as the predicted class (Y-axis). For an explicit visualization, we normalize the confusion matrix by dividing each element with the sum of each row. In addition, the kappa coefficient is leveraged to evaluate consistency and classification precision. It considers both the OA and the variations in the number of samples in each category.

[5]https://github.com/LossNAN/I3D-Tensorflow
[6]https://github.com/metalbubble/TRN-pytorch
[7]https://github.com/facebookresearch/SlowFast
[8]https://github.com/facebookresearch/SlowFast/tree/master/projects/multigrid

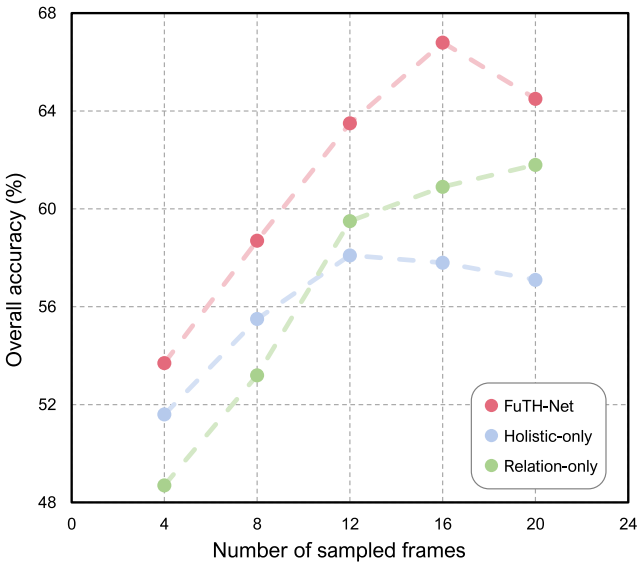


Fig. 7. FuTH-Net versus Holistic-only versus Relation-only. OAs of FuTH-Net (red), Holistic-only (blue), and Relation-only (green) with different numbers of sampled frames on the ERA dataset.

TABLE III
COMPARISON WITH DIFFERENT HYBRID MODELS. WE COMPARE OUR FUTH-NET WITH DIFFERENT HYBRID MODELS USING DIFFERENT FUSION METHODS ON THE ERA AND DRONE-ACTION DATASETS

Model	ERA Dataset		Drone-Action Dataset	
	Concatenation	Ours	Concatenation	Ours
C3D+TRN	46.7	45.3	58.4	60.1
P3D+TRN	51.9	52.4	82.9	86.6
I3D+TRN	58.4	60.8	84.3	85.2
FuTH-Net	**64.8**	**66.8**	**87.7**	**88.4**

B. Ablation Studies

To evaluate the complementarity between two pathways and effectiveness of the fusion module, we conduct ablation studies on the ERA and Drone-Action datasets.

1) Complementarity: We investigate the complementarity by comparing our FuTH-Net with its single-pathway versions on the EAR dataset. Specifically, instead of simultaneously utilizing both pathways, Holistic-only and Relation-only make use of holistic representation and temporal relation pathways, respectively. For a comprehensive study, we compare these models under variant video sampling strategies. As shown in Fig. 7, we sample 4, 8, 12, 16, and 20 frames from each

TABLE IV

COMPARISON WITH STATE-OF-THE-ART METHODS ON THE DRONE-ACTION DATASET. WE SHOW THE PER-CLASS PRECISION AND OA ON THE TEST SET. THE BEST PRECISION/ACCURACY IS SHOWN IN BOLD

Model	kicking	walking front/back	running side	jogging side	walking side	hitting stick	running front/back	stabbing	jogging front/back	clapping	hitting bottle	boxing	waving hands	OA	κ
C3D†	48.3	61.5	23.5	77.3	12.0	31.0	0.71	0.34	47.4	28.6	24.1	29.6	00.0	31.6	0.25
C3D‡	31.0	80.8	35.3	13.6	48.0	24.1	21.4	0.69	42.1	42.9	37.9	0.37	00.0	30.3	0.24
P3D†-ResNet-199	**100**	73.1	41.2	86.4	96.0	**100**	57.1	93.1	36.8	78.6	93.1	85.2	**100**	83.0	0.81
P3D‡-ResNet-199	**100**	69.2	47.1	72.7	96.0	**100**	50.0	82.8	52.6	85.7	93.1	88.9	**100**	82.3	0.81
I3D†-Inception-v1	78.3	70.4	28.6	17.9	84.2	15.2	47.1	13.6	50.0	75.0	64.1	60.0	60.0	50.7	0.79
I3D‡-Inception-v1	**100**	88.9	42.1	70.6	**100**	**100**	**64.3**	**100**	20.0	90.9	**100**	93.8	**100**	85.5	0.84
TRN†-BNInception	96.6	61.5	41.2	68.2	96.0	**100**	35.7	96.6	47.4	**100**	82.8	88.9	**100**	80.6	0.83
TRN‡-Inception-v3	**100**	**96.2**	52.9	86.4	**100**	**100**	28.6	89.7	42.1	**100**	86.2	85.2	**100**	85.0	0.82
SlowFast‡	**100.0**	88.5	52.9	95.5	92.0	**100.0**	57.1	89.7	**68.4**	92.9	86.2	**96.3**	71.4	86.7	0.86
Multigrid‡	93.1	92.3	47.1	**100.0**	**100.0**	86.2	50.0	**100.0**	63.2	**100.0**	82.8	88.9	92.9	86.4	0.85
FuTH-Net	**100**	**96.2**	**58.8**	90.9	**100**	**100**	28.6	96.6	52.6	**100**	89.7	**96.3**	**100**	**88.4**	**0.87**

TABLE V

ABLATION STUDIES OF THE FUSION MODULE ON THE ERA AND DRONE-ACTION DATASETS. WE SHOW THE OAS OF FUTH-NET AND FUTH-CONCAT AND COMPARE THEM WITH HOLISTIC-ONLY AND RELATION-ONLY NETWORKS. THE BEST ACCURACIES ARE SHOWN IN BOLD

Model	fusion	ERA Dataset	Drone-Action
Holistic-only[1]	-	57.3	84.6
Relation-only[2]	-	60.4	85.0
FuTH-Max	Max	60.1	81.4
FuTH-Average	Average	61.9	82.6
FuTH-Concat	Concatenation	64.8	85.4
FuTH-Bilinear	Bilinear	63.2	82.5
FuTH-Sum	Sum	64.7	84.8
FuTH-2DConv	2D conv	65.1	86.6
FuTH-3DConv	3D conv	65.7	87.2
FuTH-Net	Ours	**66.8**	**88.4**

[1] Holistic-only is the network with only the holistic representation pathway on top of the backbone.

[2] Relation-only is the network with only the temporal relation pathway on top of the backbone.

TABLE VI

ABLATION STUDIES ON GENERATIONS OF THE FUSED FEATURE z. WE SHOW THE OAS OF MODELS WITH DIFFERENT ADDITIONAL FEATURES ON THE ERA AND DRONE-ACTION DATASETS. THE BEST ACCURACIES ARE SHOWN IN BOLD

Additional feature	ERA Dataset	Drone-Action
None	66.0	87.3
Temporal relation $\boldsymbol{l}$	66.2	87.0
Holitic feature $\boldsymbol{g}$	**66.8**	**88.4**

video clip and show OAs. It can be observed that FuTH-Net exhibits superior performance than the other two competitors under all sampling strategies. The combination of the two pathways brings in significant improvements, demonstrating that the multiscale temporal dependencies captured by the temporal relation pathway are largely complementary with the holistic feature.

Moreover, we note that Holistic-only outperforms Relation-only when four or eight frames are used but is surpassed by Relation-only with increasing frames. The reason could be that a few frames are not enough for the learning of multiscale temporal relations. Another interesting observation is that the performance of Holistic-only deteriorates when the number of sampled frames is larger than 12, which might result from information redundancy. This also has a negative effect on FuTH-Net and brings a decrement of 2.3% with the number of sampled frames increasing from 16 to 20. Finally, FuTH-Net reaches the best performance at 16 frames.

In addition, we jointly leverage holistic spatiotemporal features and multiscale temporal relations for video classification. For validating the effectiveness of this combination, we compare our model with other hybrid models (i.e., C3D + TRN, P3D + TRN, and I3D + TRN) on two datasets using two fusion methods, concatenation and our fusion module. The numerical results are reported in Table III. We can observe that compared to other hybrid models, our FuTH-Net achieves the best performance with different fusion methods on two datasets. Moreover, we note that hybrid models with our fusion module outperform those with concatenation in general. Another interesting observation is that the three hybrid models do not achieve better performance than single models (i.e., TRN). For example, I3D + TRN with our fusion module achieves an OA of 60.8%, while TRN‡-Inception-v3 obtains an OA of 64.3%.

2) Fusion Module: As an important component in our framework, the fusion module aims to integrate features from both pathways. To validate its effectiveness, we compare

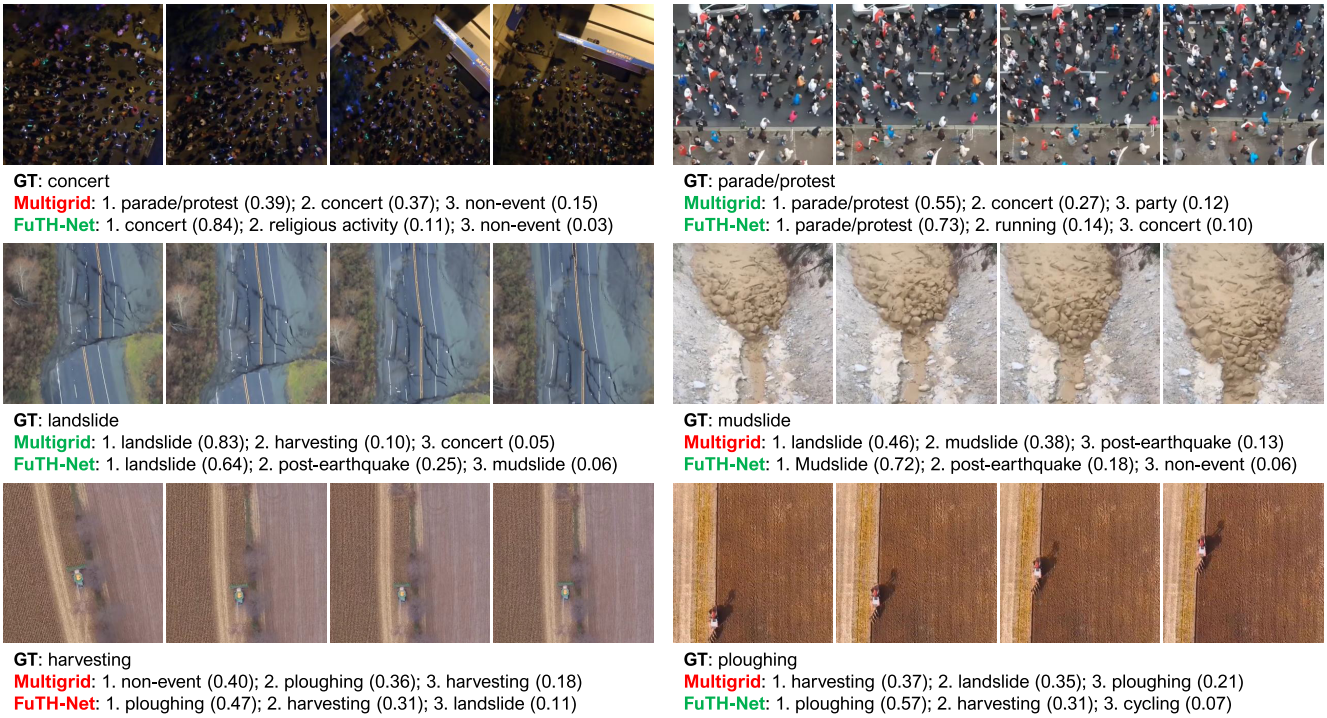
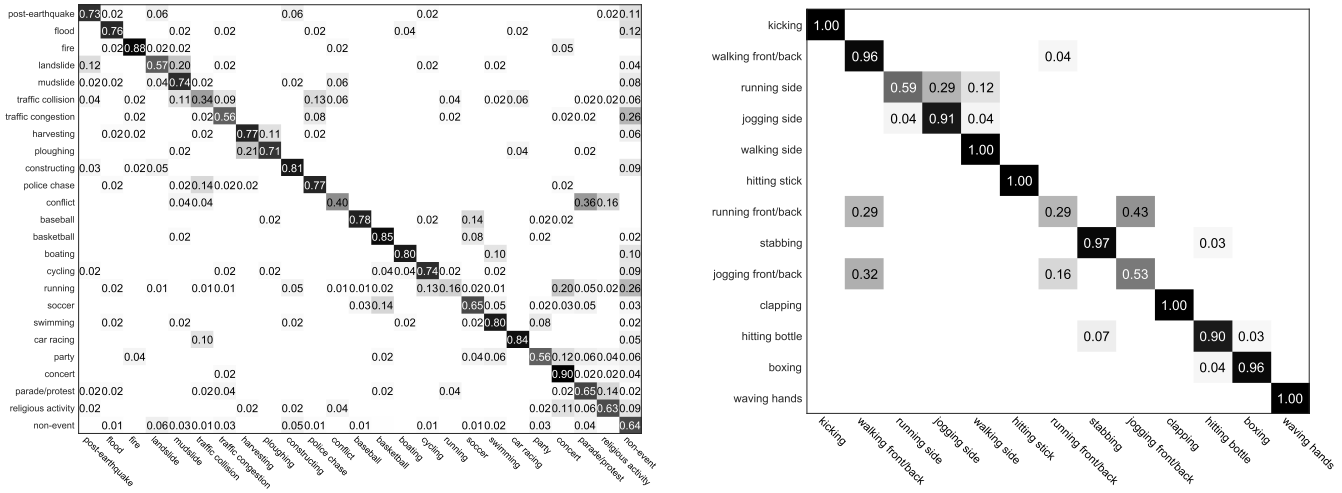


Fig. 8. Examples of predicted results on the ERA dataset. We show the results of the second best architecture, TRN, and our FuTH-Net. The ground-truth label and top three predictions of each model are reported. Four frames are selected with 1-s interval from each example video.



(a) (b)

Fig. 9. Confusion matrices of the proposed network. (a) ERA dataset. (b) Drone-Action dataset.

the fusion module with several commonly used integration operation, such as max, average, concatenation, bilinear, sum, 2-D conv, and 3-D conv. Notably, for 2-D and 3-D convs, the input is the concatenation of feature maps from the last convolutional layers of two pathways. Table V compares FuTH-Net to other models with different fusion modules on both the ERA and Drone-Action datasets. As can be seen in this table, FuTH-Net provides better results than models with other different fusion methods and models with single pathways, which demonstrates that our fusion module can effectively encode high-level interactions between the two features and improve the performance.

Moreover, we concatenate an additional holistic feature $\boldsymbol{g}$ with the modulated feature to yield the final fused feature $\boldsymbol{z}$. For ablating this design, we concatenate different additional features, i.e., none and temporal relation $\boldsymbol{l}$, with the modulated feature to obtain the final fused feature $\boldsymbol{z}$. We use these additional features to conduct ablation studies on different generations of the fused feature $\boldsymbol{z}$. The numerical results are reported in Table VI. We can observe that the model with

Fig. 10. Examples of predicted results on the Drone-Action dataset. We show the results of the second best architecture, TRN, and our FuTH-Net. The ground-truth label and top three predictions of each model are reported. Four frames are selected with 1-s interval from each example video.

holistic feature $\boldsymbol{g}$ as the additional feature outperforms other models. Richer spatial information introduced from the holistic feature can improve the discriminant ability for events with simple dynamics.

C. Results on the ERA Dataset

We compare the proposed FuTH-Net and other competitors on the ERA dataset and report the numerical results in Table II. As we can see, our model has a superb performance and provides an OA of 66.8%, which is 1.5% higher than the second best model, Multigrid[‡]. Also, our model and Multigrid achieve the same best kappa coefficient (0.63). In addition, the per-class precision is also reported to evaluate the performance of different models on each class. In particular, our model achieves the highest per-class precisions for some challenging categories, such as concert (89.8%), car racing (84.2%), and parade/protest (65.3%). This is mainly because our FuTH-Net is able to capture complex dynamic information, which is crucial to distinguish events with insignificant interclass variances. Taking concert and parade/protest (cf. the first row of Fig. 8) for example, they have something in common (e.g., crowd and street). However, the temporal dynamics of crowds in these two events are very different (concert: moving randomly or standing still and parade/protest: moving toward a certain direction). We can see that our FuTH-Net correctly predicts these two events. This can also be seen from Table II that our network gains the highest precisions for these two classes, showing its effectiveness for temporal relational reasoning.

Moreover, the performance on class nonevent can reflect whether a model can distinguish specific events from normal videos. Notably, our model produces the best precision (63.9%) for nonevent, which illustrates that our method is able to capture discriminative spatiotemporal features for inferring the existence of events.

Finally, the confusion matrix in Fig. 9(a) shows more details. We can observe that some events, including similar objects and scenes (e.g., "landslide versus mudslide," "traffic collision versus police chase," "harvesting versus plowing," and "concert versus party"), tend to be misclassified. Other competitors also suffer from this problem. Fig. 8 shows some predictions of FuTH-Net and the second best model (i.e., Multigrid). It can be observed that there are a lot of visual similarities existing in textures, objects, and scenes of these events.

D. Results on the Drone-Action Dataset

This section compares FuTH-Net and state-of-the-art methods on the Drone-Action dataset, and quantitative results are reported in Table IV. Our FuTH-Net achieves the highest OA, 88.4%, and compared to SlowFast that is the second best model, an increment of 1.7% can be obtained. Moreover, our model achieves the best kappa coefficient (0.87).

Besides, it is interesting to note that FuTH-Net shows good performance in recognizing actions in which effectively sensing motion speeds is crucial for a successful prediction. For instance, the proposed network gains the highest precisions for walking side (100.0%), running side (58.8%), and jogging side (100.0%). To further illustrate this, we show some predictions of FuTH-Net and the second best model (i.e., SlowFast) in Fig. 10. As can be observed, the motion speeds of walking side and running side are variant, and our FuTH-Net succeeds

in identifying them with high confidence. The bottom right example shows that running front/back is misclassified by both FuTH-Net and SlowFast since human poses and motion speeds are very similar in this angle of view. Furthermore, the confusion matrix of the proposed network on the Drone-Action dataset shown in Fig. 9(b) also suggests that running front/back is easily misidentified as jogging front/back.

IV. Conclusion

In this article, a novel method is proposed to learn feature representations from aerial videos using a two-pathway network, termed FuTH-Net. Specifically, the proposed network exploits inflated 3-D conclusions to capture a holistic feature on a holistic representation pathway. Simultaneously, a temporal relation block learns temporal relations across multiple frames on a temporal relation pathway. A novel fusion module is applied to fuse outputs from the two pathways for producing a more discriminative video representation. Furthermore, we conduct extensive experiments on two aerial video recognition datasets, ERA and Drone-Action. On the one hand, we perform ablation studies to validate the complementarity between the two pathways as well as the effectiveness of the proposed fusion module. On the other hand, we compare our model with other state-of-the-art methods. Experimental results demonstrate that the introduction of the temporal relation pathway can enhance the ability of capturing representative temporal relations. Besides, our fusion module is capable of learning high-level interactions between the holistic features and temporal relations to further boost the performance. The outstanding performance on the two datasets further illustrates the superior capability of FuTH-Net for remote sensing video recognition and its powerful generalization capability across different tasks (event classification and human action recognition).

References

[1] G. Pajares, "Overview and current status of remote sensing applications based on unmanned aerial vehicles (UAVs)," *Photogramm. Eng. Remote Sens.*, vol. 81, no. 4, pp. 281–330, 2015.

[2] T. Xiang, G.-S. Xia, and L. Zhang, "Mini-unmanned aerial vehicle-based remote sensing: Techniques, applications, and prospects," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 3, pp. 29–63, Sep. 2019.

[3] X. X. Zhu *et al.*, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[4] I. Colomina and P. Molina, "Unmanned aerial systems for photogrammetry and remote sensing: A review," *ISPRS J. Photogram. Remote Sens.*, vol. 92, pp. 79–97, Jun. 2014.

[5] N. Tijtgat, W. Van Ranst, T. Goedeme, B. Volckaert, and F. De Turck, "Embedded real-time object detection for a UAV warning system," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Oct. 2017, pp. 2110–2118.

[6] M. Teutsch and W. Kruger, "Detection, segmentation, and tracking of moving objects in UAV videos," in *Proc. IEEE 9th Int. Conf. Adv. Video Signal-Based Surveill. (AVSS)*, Sep. 2012, pp. 313–318.

[7] S. Zhang, "Object tracking in unmanned aerial vehicle (UAV) videos using a combined approach," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2005, pp. 1–4.

[8] S. Xuan, S. Li, M. Han, X. Wan, and G. S. Xia, "Object tracking in satellite videos by improved correlation filters with motion estimations," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 2, pp. 1074–1086, Feb. 2020.

[9] Q. Li, L. Mou, Q. Xu, Y. Zhang, and X. X. Zhu, "R^3-Net: A deep network for multi-oriented vehicle detection in aerial images and videos," 2018, *arXiv:1808.05560*.

[10] V. N. Dobrokhodov, I. I. Kaminer, K. D. Jones, and R. Ghabcheloo, "Vision-based tracking and motion estimation for moving targets using small UAVs," in *Proc. Amer. Control Conf. (ACC)*, 2006, pp. 1–32.

[11] L. Wang, F. Chen, and H. Yin, "Detecting and tracking vehicles in traffic by unmanned aerial vehicles," *Automat. Construct.*, vol. 72, pp. 294–308, Dec. 2016.

[12] A. Puri, "A survey of unmanned aerial vehicles (UAV) for traffic surveillance," Dept. Comput. Sci. Eng., Univ. South Florida, Tampa, FL, USA, Tech. Rep., 2005, pp. 1–29. [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.108.8384&rep=rep1&type=pdf

[13] K. Kanistras, G. Martins, M. J. Rutherford, and K. P. Valavanis, "A survey of unmanned aerial vehicles (UAVs) for traffic monitoring," in *Proc. Int. Conf. Unmanned Aircr. Syst. (ICUAS)*, May 2013, pp. 221–234.

[14] A. Puri, K. P. Valavanis, and M. Kontitsis, "Statistical profile generation for traffic monitoring using real-time UAV based video data," in *Proc. Medit. Conf. Control Autom.*, Jun. 2007, pp. 1–6.

[15] Q. Feng, J. Liu, and J. Gong, "UAV remote sensing for urban vegetation mapping using random forest and texture analysis," *Remote Sens.*, vol. 7, no. 1, pp. 1074–1094, Jan. 2015.

[16] J. Everaerts, "The use of unmanned aerial vehicles (UAVs) for remote sensing and mapping," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 37, pp. 1187–1192, Jul. 2008.

[17] A. Laliberte, "Unmanned aerial vehicle-based remote sensing for rangeland assessment, monitoring, and management," *J. Appl. Remote Sens.*, vol. 3, no. 1, Aug. 2009, Art. no. 033542.

[18] L. Mou, Y. Hua, P. Jin, and X. X. Zhu, "ERA: A dataset and deep learning benchmark for event recognition in aerial videos," 2020, *arXiv:2001.11394*.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[20] R. Ranjan *et al.*, "Deep learning for understanding faces: Machines may be just as good, or better, than humans," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 66–83, Jan. 2018.

[21] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1725–1732.

[22] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2625–2634.

[23] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4694–4702.

[24] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2014, pp. 1–9.

[25] Z. Qiu, T. Yao, and T. Mei, "Learning deep spatio-temporal dependence for semantic video segmentation," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 939–949, Apr. 2018.

[26] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–8.

[27] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *Int. J. Comput. Vis.*, vol. 79, no. 3, pp. 299–318, 2008.

[28] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 3551–3558.

[29] X. Wang, A. Farhadi, and A. Gupta, "Actions~ transformations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2658–2667.

[30] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1933–1941.

[31] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[32] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[33] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler, "Convolutional learning of spatio-temporal features," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2010, pp. 140–153.

[34] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1510–1517, Jun. 2018.

[35] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the Kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6299–6308.

[36] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5533–5541.

[37] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 6450–6459.

[38] D. Tran, H. Wang, M. Feiszli, and L. Torresani, "Video classification with channel-separated convolutional networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5552–5561.

[39] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[40] J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7083–7093.

[41] L. Mou and X. X. Zhu, "Spatiotemporal scene interpretation of space videos via deep neural network and tracklet analysis," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 1823–1826.

[42] X. Liu, J.-Y. Lee, and H. Jin, "Learning video representations from correspondence proposals," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4273–4281.

[43] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 803–818.

[44] L. Wang *et al.*, "Temporal segment networks for action recognition in videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2740–2755, Nov. 2019.

[45] T. Shu, D. Xie, B. Rothrock, S. Todorovic, and S.-C. Zhu, "Joint inference of groups, events and human roles in aerial videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4576–4584.

[46] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6202–6211.

[47] C.-Y. Wu, R. Girshick, K. He, C. Feichtenhofer, and P. Krahenbuhl, "A multigrid method for efficiently training video models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 153–162.

[48] Y. Lyu, G. Vosselman, G.-S. Xia, A. Yilmaz, and M. Y. Yang, "UAVid: A semantic segmentation dataset for UAV imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 165, pp. 108–119, Jul. 2020.

[49] L. Mou and X. X. Zhu, "Vehicle instance segmentation from aerial image and video using a multitask learning residual fully convolutional network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6699–6711, Nov. 2018.

[50] Y. Lyu, G. Vosselman, G.-S. Xia, and M. Y. Yang, "LIP: Learning instance propagation for video object segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1–10.

[51] F. Yang, G.-S. Xia, G. Liu, L. Zhang, and X. Huang, "Dynamic texture recognition by aggregating spatial and temporal features via ensemble SVMs," *Neurocomputing*, vol. 173, pp. 1310–1321, Jan. 2016.

[52] J. Gao, Q. Wang, and X. Li, "PCC Net: Perspective crowd counting via spatial convolutional network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3486–3498, Oct. 2020.

[53] X. Li, Z. Zhao, and Q. Wang, "ABSSNet: Attention-based spatial segmentation network for traffic scene understanding," *IEEE Trans. Cybern.*, early access, Feb. 2, 2021, doi: 10.1109/TCYB.2021.3050558.

[54] C. Feichtenhofer, A. Pinz, and R. Wildes, "Spatiotemporal residual networks for video action recognition," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 29, 2016, pp. 3468–3476.

[55] M. W. Gardner and S. R. Dorling, "Artificial neural networks (the multilayer perceptron)—A review of applications in the atmospheric sciences," *Atmos. Environ.*, vol. 32, nos. 14–15, pp. 2627–2636, 1998.

[56] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*.

[57] P. Su *et al.*, "Adapting object detectors with conditional domain normalization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 403–419.

[58] D. He *et al.*, "StNet: Local and global spatial-temporal modeling for action recognition," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2019, pp. 8401–8408.

[59] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[60] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. Int. Conf. Comput. Statist. (COMPSTAT)*, 2010, pp. 177–186.

[61] A. G. Perera, Y. W. Law, and J. Chahl, "Drone-Action: An outdoor recorded drone video dataset for action recognition," *Drones*, vol. 3, no. 4, p. 82, Nov. 2019.

[62] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*.

[63] W. Kay *et al.*, "The Kinetics human action video dataset," 2017, *arXiv:1705.06950*.

[64] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, "A short note about Kinetics-600," 2018, *arXiv:1808.01340*.

[65] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 448–456.

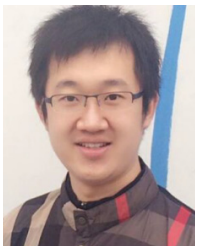
[66] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[67] R. Goyal *et al.*, "The 'something something' video database for learning and evaluating visual common sense," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5842–5850.

[68] M. Monfort *et al.*, "Moments in time dataset: One million videos for event understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 502–508, Feb. 2020.

Pu Jin (Member, IEEE) received the bachelor's degree in electronic information science and technology from Wuhan University, Wuhan, China, in 2017, the master's degree in Earth oriented space science and technology (ESPACE) from the Technical University of Munich (TUM), Munich, Germany, in 2020, and the master's degree in photogrammetry and remote sensing from Wuhan University in 2021. He is currently pursuing the Ph.D. degree with the German Aerospace Center (DLR), Weßling, Germany, and TUM.

His research interests include remote sensing, computer vision, and deep learning, especially their applications in remote sensing.



Lichao Mou received the bachelor's degree in automation from the Xi'an University of Posts and Telecommunications, Xi'an, China, in 2012, the master's degree in signal and information processing from the University of Chinese Academy of Sciences (UCAS), Beijing, China, in 2015, and the Dr.Ing. degree from the Technical University of Munich (TUM), Munich, Germany, in 2020.

In 2015, he spent six months at the Computer Vision Group, University of Freiburg, Freiburg im Breisgau, Germany. In 2019, he was a Visiting Researcher with the Cambridge Image Analysis Group (CIA), University of Cambridge, Cambridge, U.K. Since 2019, he has been a Research Scientist with DLR-IMF and an AI Consultant with the Helmholtz Artificial Intelligence Cooperation Unit (HAICU). He is currently a Guest Professor with the Munich AI Future Lab AI4EO, TUM, and the Head of the Visual Learning and Reasoning Team, Department "EO Data Science," Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Weßling, Germany.

Dr. Mou was a recipient of the First Place in the 2016 IEEE GRSS Data Fusion Contest and finalists for the Best Student Paper Award at the 2017 Joint Urban Remote Sensing Event and the 2019 Joint Urban Remote Sensing Event.

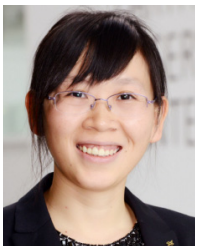
Yuansheng Hua (Graduate Student Member, IEEE) received the bachelor's degree in remote sensing science and technology from Wuhan University, Wuhan, China, in 2014, the master's degree in Earth-oriented space science and technology (ESPACE) from the Technical University of Munich (TUM), Munich, Germany, in 2018, and the master's degree in photogrammetry and remote sensing from Wuhan University in 2019. He is currently pursuing the Ph.D. degree with the German Aerospace Center (DLR), Weßling, Germany, and TUM.

In 2019, he was a Visiting Researcher with Wageningen University and Research, Wageningen, The Netherlands. His research interests include remote sensing, computer vision, and deep learning, especially their applications in remote sensing.

Gui-Song Xia (Senior Member, IEEE) received the Ph.D. degree in image processing and computer vision from CNRS LTCI, Télécom ParisTech, Paris, France, in 2011.

From 2011 to 2012, he has been a Post-Doctoral Researcher with the Centre de Recherche en Mathématiques de la Decision, CNRS, Paris-Dauphine University, Paris, for one and half years. He is currently working as a Full Professor with Wuhan University, Wuhan, China, leading a group working toward computer vision and photogrammetry. He was a Visiting Scholar at DMA, École Normale Supérieure (ENS-Paris), Paris, for two months in 2018. His research interests include mathematical modeling of images and videos, structure from motion, perceptual grouping, and remote sensing image understanding.

Dr. Xia serves on the Editorial Board of several journals, including *ISPRS Journal of Photogrammetry and Remote Sensing*, *Pattern Recognition*, *Signal Processing: Image Communication*, *Journal of Remote Sensing*, and *Frontiers in Computer Science: Computer Vision*.



Xiao Xiang Zhu (Fellow, IEEE) received the M.Sc. degree, the Dr.Ing. degree, and the Habilitation degree in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

Since 2019, she has been a Co-Coordinator of the Munich Data Science Research School. Since 2019, she has been heading the Helmholtz Artificial Intelligence—Research Field "Aeronautics, Space and Transport." Since May 2020, she has been the Director of the international future AI lab "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond," Munich. Since October 2020, she has been the Co-Director of the Munich Data Science Institute (MDSI), TUM. She was a Guest Scientist or a Visiting Professor with the Italian National Research Council (CNR-IREA), Naples, Italy, in 2009; Fudan University, Shanghai, China, in 2014; The University of Tokyo, Tokyo, Japan, in 2015; and the University of California at Los Angeles, Los Angeles, CA, USA, in 2016. She is currently a Professor with the Data Science in Earth Observation (formerly Signal Processing in Earth Observation), TUM, and the Head of the Department "EO Data Science," Remote Sensing Technology Institute, German Aerospace Center (DLR), Weßling, Germany. She is also a Visiting AI Professor with ESA's Phi-lab. Her main research interests are remote sensing and Earth observation, signal processing, machine learning, and data science, with a special application focus on global urban mapping.

Dr. Zhu is a member of the Young Academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities, the German National Academy of Sciences Leopoldina, and the Bavarian Academy of Sciences and Humanities. She serves on the scientific advisory board in several research organizations, including the German Research Center for Geosciences (GFZ) and Potsdam Institute for Climate Impact Research (PIK). She is an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. She serves as an Area Editor for the Special Issues of *IEEE Signal Processing Magazine*.