2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 978-1-6654-4509-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/CVPR46437.2021.01600

UAV-Human: A Large Benchmark for Human Behavior Understanding with Unmanned Aerial Vehicles

Tianjiao Li[1] Jun Liu[1] * Wei Zhang[2] * Yun Ni[1] Wenqian Wang[2] Zhiheng Li[2]

[1] Information Systems Technology and Design, Singapore University of Technology and Design, Singapore

[2] School of Control Science and Engineering, Shandong University, Jinan, Shandong

tianjiao_li@mymail.sutd.edu.sg, {jun_liu, ni_yun}@sutd.edu.sg

davidzhang@sdu.edu.cn, {wqwang, zhihengli}@mail.sdu.edu.cn

Abstract

Human behavior understanding with unmanned aerial vehicles (UAVs) is of great significance for a wide range of applications, which simultaneously brings an urgent demand of large, challenging, and comprehensive benchmarks for the development and evaluation of UAV-based models. However, existing benchmarks have limitations in terms of the amount of captured data, types of data modalities, categories of provided tasks, and diversities of subjects and environments. Here we propose a new benchmark - UAV-Human - for human behavior understanding with UAVs, which contains 67,428 multi-modal video sequences and 119 subjects for action recognition, 22,476 frames for pose estimation, 41,290 frames and 1,144 identities for person re-identification, and 22,263 frames for attribute recognition. Our dataset was collected by a flying UAV in multiple urban and rural districts in both daytime and nighttime over three months, hence covering extensive diversities w.r.t subjects, backgrounds, illuminations, weathers, occlusions, camera motions, and UAV flying attitudes. Such a comprehensive and challenging benchmark shall be able to promote the research of UAV-based human behavior understanding, including action recognition, pose estimation, re-identification, and attribute recognition. Furthermore, we propose a fisheye-based action recognition method that mitigates the distortions in fisheye videos via learning unbounded transformations guided by flat RGB videos. Experiments show the efficacy of our method on the UAV-Human dataset.

1. Introduction

Given the flexibility and capability of long-range tracking, unmanned aerial vehicles (UAVs) equipped with cameras are often used to collect information from remote for the scenarios where it is either impossible or not sensible to use ground cameras [24, 35, 22]. One particular area where UAVs are often deployed is human behavior understanding and surveillance in the wild, where video sequences of human subjects can be collected for analysis (such as action recognition, pose estimation, human re-identification, and attribute analysis), and for subsequent decision making.

Compared to videos collected by common ground cameras, the video sequences captured by UAVs generally present more diversified yet unique viewpoints, more obvious motion blurs, and more varying resolutions of the subjects, owing to the fast motion and continuously changing attitudes and heights of the UAVs during flight. These factors lead to significant challenges in UAV-based human behavior understanding, clearly requiring the design and development of human behavior understanding methods specifically taking the unique characteristics of UAV application scenarios into consideration.

Existing works [8] have demonstrated the great importance of leveraging large, comprehensive, and challenging benchmarks to develop and evaluate the state-of-the-art deep learning methods for handling various computer vision tasks. However, in the UAV-based human behavior understanding area, existing datasets [3, 26] have limitations in multiple aspects, including: (1) Very limited number of samples, while a large scale of the dataset is often important for mitigating over-fitting issues and enhancing the generalization capability of the models developed on it. (2) Limited number and limited diversity of subjects, while the diversities of human ages, genders, and clothing are crucial for developing robust models for analyzing the behaviors of various subjects. (3) Constrained capturing conditions. In practical application scenarios, UAVs often need to work in various regions (e.g., urban, rural, and even mountain and river areas), under diversified weathers (e.g., windy and rainy weathers), in different time periods (e.g, daytime and nighttime). However, samples in existing datasets are usu-

*Corresponding Authors

978-1-6654-4509-2/21/$31.00 ©2021 IEEE
DOI 10.1109/CVPR46437.2021.01600

ally collected under similar conditions, obviously simplifying the challenges in real-world UAV application scenarios. (4) Limited UAV viewpoints and flying attitudes. UAVs can experience frequent (and sometimes abrupt) position shifts during flying, which not only cause obvious viewpoint variations and motion blurs, but also lead to significant resolution changes. However, the UAVs in most of the existing datasets only present slow and slight motions with limited flying attitude variations. (5) Limited types of data modalities. In practical scenarios, we often need to deploy different types of sensors to collect data under different conditions. For example, infrared (IR) sensors can be deployed on UAVs for human search and rescue in the nighttime, while fisheye cameras are often used to capture a broad area. This indicates the significance of collecting different data modalities, to facilitate the development of various models for analyzing human behaviors under different conditions. However, most of the existing UAV datasets provide conventional RGB video samples only. (6) Limited categories of provided tasks and annotations. As for UAV-based human behavior understanding, various tasks, such as action recognition, pose estimation, re-identification (ID), and attribute analysis, are all of great significance, which indicates the importance of providing thorough annotations of various tasks for a comprehensive behavior analysis. However, most of the existing datasets provide annotations for one or two tasks only.

The aforementioned limitations in existing datasets clearly show the demand of a larger, more challenging, and more comprehensive dataset for human behavior analysis with UAVs. Motivated by this, in this work, we create UAV-Human, the first large-scale multi-modal benchmark in this domain. To construct this benchmark, we collect samples by flying UAVs equipped with multiple sensors in both daytime and nighttime, over three different months, and across multiple rural districts and cities, which thus brings a large number of video samples covering extensive diversities w.r.t human subjects, data modalities, capturing environments, and UAV flying attitudes and speeds, etc.

Specifically, a total of 22,476×3 video sequences (consisting of three sensors: Azure DK, fisheye camera and night-vision camera) with 119 distinct subjects and 155 different activity categories are collected for action recognition; 22,476 frames with 17 major keypoints are annotated for pose estimation; 41,290 frames with 1,144 identities are collected for person re-identification; and 22,263 frames with 7 distinct characteristics are labelled for attribute recognition, where the captured subjects present a wide range of ages (from 7 to 70) and clothing styles (from summer dressing to fall dressing). Meanwhile the capturing environments contain diverse scenes (45 sites, including forests, riversides, mountains, farmlands, streets, gyms, and buildings), different weather conditions (sunny, cloudy, rainy, and windy), and various illumination conditions (dark and bright). Besides, different types of UAV flying attitudes, speeds, and trajectories are adopted to collect data, and thus our dataset covers very diversified yet practical viewpoints and camera motions in UAV application scenarios. Furthermore, the equipped different sensors enable our dataset to provide rich data modalities including RGB, depth, IR, fisheye, night-vision, and skeleton sequences.

Besides introducing the UAV-Human dataset, in this paper, we also propose a method for action recognition in fisheye UAV videos. Thanks to the wide angle of view, fisheye cameras can capture a large area in one shot and thus are often deployed on UAVs for surveillance. However, the provided wide angle in turn brings large distortions into the collected videos, making fisheye-based action recognition quite challenging. To mitigate this problem, we design a Guided Transformer I3D model to learn an unbounded transformation for fisheye videos under the guidance of flat RGB images. Experimental results show that such a design is able to boost the performance of action recognition using fisheye cameras.

2. Related Work

In this section, we review the previous UAV-based human behavior datasets that are relevant to our dataset. Besides, since there is a lack of *large UAV-based* datasets for multi-modal behavior analysis, we also review some *ground camera-based* multi-modal datasets.

UAV-Based Human Behavior Understanding Datasets. Thanks to the flexibility, UAVs have been used in many scenarios where ground cameras may be difficult to be deployed, and some UAV-based benchmarks [26, 3, 26, 39, 1, 2, 25, 22] have been introduced for human behavior understanding. However, to the best of our knowledge, all the existing benchmarks have limitations with regard to the dataset size, the diversities of scenes, the provided task categories, and captured data modality types, etc.

Okutama-Action [3] is a relatively small dataset for human action recognition, in which the RGB videos were collected over a baseball field with UAVs. This dataset only includes 12 action classes performed by 9 subjects. The small number of video samples and subjects, and the relatively simple scene obviously hinder its application for more challenging real-world scenarios.

UAV-Gesture [26] is a dataset collected for UAV control gesture and pose analysis. This dataset provides 119 RGB video samples for 13 UAV control gestures, that were performed by 10 subjects with relatively monotonous backgrounds.

PRAI-1581 [39] is a UAV-based dataset for person ReID. However this dataset only provides a single RGB modality and annotations for a single ReID task.

AVI [31] is a human behavior understanding dataset for violent action recognition and pose estimation. Although this dataset provides annotations for two tasks, this dataset is quite small and lacks diversities in multiple aspects. It contains 2K RGB images only. Besides, only 5 action classes were performed by 25 subjects with small age differences (18 to 25).

Compared to all the existing UAV human behavior analysis datasets, our UAV-Human has significant advantages and provides many more videos and images, many more actions and poses, many more various scenes and backgrounds, much more diversified viewpoints and flying attitudes, many more data modalities, much more rich annotations, and many more tasks, etc. Thus our dataset shall be able to serve as a comprehensive and challenging benchmark for human behavior analysis with UAVs.

Ground Camera-Based Multi-Modal Human Behavior Datasets. Since our Human-UAV dataset provides 6 different types of data modalities, here we also briefly review some of the ground camera-based human behavior datasets [18, 13, 19, 27, 11, 14, 15, 21, 36, 16] that contain multi-modal data. SYSU 3DHOI [11] is an indoor RGB-D dataset for human action recognition, which includes 480 video samples in 12 action categories, captured from 40 subjects. UWA3D Multiview II [27] was captured from 10 subjects with 4 fixed camera angles for cross-view action understanding, which includes 1,075 videos and 30 action classes. PKU-MMD [18] contains 1,076 video sequences recorded in an indoor environment for human action recognition. Varying-View RGB-D [13] consists of 25,600 videos for action recognition, which were captured with ground-view cameras mounted on robots in the same indoor environment. NTU RGB+D 120 [19] is a large-scale dataset captured using three fixed cameras, which provides four data modalities including RGB, depth, IR and skeleton data.

Unlike the aforementioned datasets that were captured using ground cameras, our UAV-Human is collected by flying UAVs with different speeds, heights, attitudes, and trajectories. Due to the flexibility of UAVs, our dataset provides unique viewpoints, obvious resolution variations, significant camera movements, and frequent motion blurs. Moreover, almost all the existing ground camera-based multi-modal human behavior datasets were collected under relatively simple, static, and monotonous scenes, while our dataset covers diverse outdoor and indoor scenes, different weather conditions, and various illumination and occlusion conditions, etc.

Fisheye-Based Human Behavior Analysis Methods. Delibasis *et al.* [7] proposed a deformable 3D human model to recognize the postures of a monitored person recorded by a fisheye camera. Srisamosorn *et al.* [32] introduced a histogram of oriented gradient descriptors to detect human body and head directions in fisheye videos. To handle distorted fisheye video samples, here we propose to use the flat RGB images to guide a spatial transformer layer, which boosts the performance of action recognition in fisheye videos.

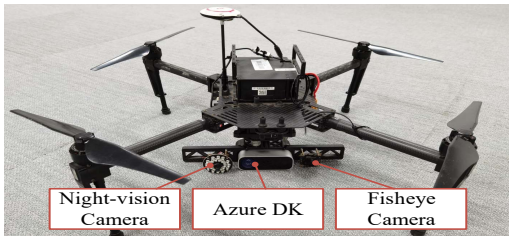


Figure 1: Illustration of our dataset collection platform deployed with multiple sensors.

3. UAV-Human Dataset

3.1. Dataset Description

We create UAV-Human, a large-scale benchmark for UAV-based human behavior understanding, which contains 67,428 annotated video sequences of 119 subjects for action recognition, 22,476 annotated frames for pose estimation, 41,290 annotated frames of 1,144 identities for person re-identification, and 22,263 annotated frames for attribute recognition. Our dataset is collected with a DJI Matrice 100 platform as shown in Figure 1. To the best of our knowledge, UAV-Human is currently the largest, most challenging, and most comprehensive UAV dataset for human action, pose, and behavior understanding. Some of the collected video samples are illustrated in Figure 2. Below we introduce the noteworthy features of our proposed dataset.

Multiple Data Modalities. We collect multiple data modalities, of fisheye videos, night-vision videos, common RGB videos, infrared (IR) sequences, and depth maps, using a UAV with multiple sensors, including a fisheye camera, a night-vision sensor, and an Azure Kinect DK as shown in Figure 1. Note that instead of using Microsoft Kinect V2 that has been widely used for indoor depth data collection [19, 28], we use the very recently released Azure Kinect DK to collect the IR and depth maps, since it is still consumer-available, yet is more powerful and more advanced, and can work better in both indoor and outdoor environments, compared to Microsoft Kinect V2. We actually observe the Azure Kinect DK works reliably and provides quite promising depth data in both indoor and outdoor scenarios during our dataset collection.

Besides the aforementioned modalities, we also fine-tune a pose estimator [9] on our dataset, considering the unique capturing viewpoints of our UAV dataset. Hence the skeleton data is also provided for each video frame.

Specifically, the RGB videos are recorded and stored at the resolution of 1920×1080. The millimeter-measured

Table 1: Comparisons among our UAV-Human dataset and some previous UAV-based human behavior analysis datasets, and ground camera-based multi-modal datasets. Our UAV-Human dataset significantly outperforms all the previous UAV-based human behavior datasets w.r.t. the data volume, the modalities, the diversities, and the labelled tasks and samples. Besides, it even obviously outperforms existing non-UAV (i.e., ground camera) based multi-modal datasets in lots of aspects.

		Ground Camera-Based Multi-Modal Datasets				**UAV-Based Datasets**				
Dataset Attribute		PKU MMD	Varying View	NTU RGBD120	Thermal World	Okutama Action	UAV Gesture	PRAI 1,581	AVI	**UAV-Human (Ours)**
Action Recog. Task	# Annotated Videos	1,076	25,600	114,480	×	43	119	×	NA	**22,476×3**
	# Annotated Subjects	66	118	106	×	9	10	×	25	**119**
	# Annotated Classes	51	40	120	×	12	13	×	5	**155**
Pose Estimation Task	# Annotated Samples	×	×	×	×	×	×	×	2,000	**22,476**
Person ReID Task	# Annotated IDs	×	×	×	516	×	×	**1,581**	×	1,144
	# Annotated Samples	×	×	×	15,118	×	×	39,461	×	**41,290**
Attribute Recog. Task	# Annotated Samples	×	×	×	×	×	×	×	×	**22,263**
Data Modality	RGB	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Depth	✓	✓	✓	×	×	×	×	×	✓
	IR	✓	×	✓	×	×	×	×	×	✓
	Joint	✓	✓	✓	×	×	✓	×	✓	✓
	Others	×	×	×	Thermal	×	×	×	×	**Fisheye Night-vision**
Sensors		Kinect V2	Kinect V2	Kinect V2	FLIR ONE	NA	GoPro 4	NA	NA	**Azure DK, Fisheye Camera, Night-vision Camera**
Capturing Scenarios	# Sites	1	1	3	3	1	1	2	1	**45**
	Indoor	✓	✓	✓	✓	×	×	×	×	✓
	Outdoor	×	×	✓	✓	✓	✓	✓	✓	✓
Challenging Weathers	Windy	×	×	×	×	×	✓	×	×	✓
	Rainy	×	×	×	✓	×	×	×	×	✓
Complex Backgrounds		×	×	✓	✓	×	×	×	×	✓
Occlusion		×	×	×	✓	✓	×	✓	×	✓
Night Scenes		×	×	×	✓	×	×	×	×	✓
Camera Views		fixed	varying	fixed	fixed	varying	varying	varying	varying	varying
UAV Attitudes	Hover	×	×	×	×	✓	✓	✓	✓	✓
	Lift	×	×	×	×	✓	×	×	×	✓
	Descent	×	×	×	×	✓	×	×	×	✓
	Cruising	×	×	×	×	×	×	✓	×	✓
	Rotating	×	×	×	×	×	×	✓	×	✓

depth maps are recorded in lossless compression formats at the resolution of 640×576. The IR sequences are also stored at the resolution of 640×576. The fisheye videos are captured by a 180°-vision camera, and the resolution is 640×480. The night-vision videos are recorded at two automatic modes, namely, color mode in the daytime and grey-scale mode in the nighttime, with the resolution at 640×480. The skeleton modality stores the positions of 17 major key-points of the human body.

The provided miscellaneous data modalities captured with different sensors that have different properties and different suitable application scenarios, shall be able to facilitate the community to exploit various methods to utilize different modalities, as well as cross-modality and modality-ensemble ones for UAV applications.

Large Variations of Capture Environments. Our UAV-Human is captured from a total of 45 different sites across multiple rural districts and cities, which thus covers various outdoor and indoor scenes, including mountain areas, forests, river and lake-side areas, farmlands, squares, streets, gyms, university campuses, and several scenes inside buildings. Such abundant variations of scenes and backgrounds bring practical challenges to the UAV-based human behavior understanding problem.

Long Time Period of Capturing. The overall time period of our dataset collection lasts for three months, across two different seasons (summer and fall), in both daytime and nighttime. Thus many distinctive features can be observed, such as the change of subject clothing fashions and surrounding styles, resulting from time period changes. In all, a long recording time period remarkably increases the diversities of the recorded video sequences.

Various Challenging Weather Conditions. During our dataset collection, we encounter many adverse weather conditions including rainy and windy. Specifically, occlusions caused by umbrellas, UAV shaking caused by strong wind, and low capture qualities caused by rain and fog, are all covered in our dataset. Consequently, extreme weather conditions obviously lead to many challenging yet practical factors for our proposed dataset.

Varying UAV Attitudes, Positions, and Views. In practical application scenarios, UAVs may have different flight attitudes, including hover, climb, descent, hovering turn, and side-ward flight, etc., which can result in significant camera shakes, affecting the locations of subjects in the captured frames, and leading to obvious motion blurs. Hence the diversified UAV flight attitudes and speeds in our dataset could encourage the community to develop robust models to handle human behavior analysis in such challenging scenarios. Besides, the hovering height (varying from 2 to 8

meters) in our dataset also brings large resolution variations of the subjects in the collected videos. Moreover, unlike many previous datasets [26], where the capturing orientations are relatively constrained, our UAV-Human dataset provides very flexible UAV views.

Multiple Human Behavior Understanding Tasks. For comprehensively analyzing human behaviors and actions from UAV videos, our proposed UAV-Human dataset provides annotations for four major tasks, namely, action recognition, pose estimation, person re-identification, and human attribute recognition, with very rich and elaborate annotations. Note that for all these tasks, besides containing the universal features introduced above, each of them has the own notable contributions that are detailed in next section.

3.2. Dataset Tasks and Annotations

Based on the aforementioned data capturing platform and settings, we collect and annotate rich samples for the following different tasks.

(1) **Action recognition with a large number of activity classes.** To collect sufficient data for this task, we invite a total of 119 distinct subjects with different ages, genders, and occupations to naturally perform various actions and poses in all captured scenes during our dataset collection, which thus enables us to collect very diversified human actions, yet still keep the action classes balanced. Specifically, we collect 155 activity classes in 6 different modalities (shown in Figure 2) covering different types of human behaviors that can be of great interest and significance for the practical UAV application scenarios. These activities include: (i) daily activities, e.g., smoking, wearing/taking off masks, (ii) productive activities, e.g., digging, fishing, mowing, cutting trees, carrying with shoulder poles, (iii) violent activities, e.g., taking a hostage, stabbing with a knife, lock picking, (iv) social interaction behaviors, e.g., walking together closely, whispering, (v) life-saving activities, e.g., calling for help, escaping, and (vi) UAV control gestures. The aforementioned action classes can naturally occur in the wild, where CCTV surveillance cameras may be unavailable, while the UAVs can be used to flexibly track the performers of such activities in these scenarios.

(2) **Pose estimation with manually-labeled human key-points.** We invite 15 volunteers to label human poses in a total of 22,476 images that are sampled from the collected videos. For each image sample, a total of 17 major body joints are manually labelled, as illustrated in Figure 3. Note that the human pose estimation task in our dataset is quite challenging, owing to the distinct UAV viewpoints, different subject resolutions, diverse backgrounds, and various illumination, weather, and occlusion conditions.

(3) **Attribute recognition with rich individual characteristics.** The proposed dataset also provides 22,263 person images for human attribute recognition. We label 7 groups of attributes, including gender, hat, backpack, upper clothing color and style, as well as lower clothing color and style as shown in Figure 4. It is worth noting that the unconstrained views recorded from the UAV platform can cause large occlusions, leading to very challenging attribute recognition scenarios.

Figure 2: Examples of action videos in our UAV-Human dataset. The first row shows different data modalities. The second and third rows show two video sequences of significant camera motions and view variations, caused by continuously varying flight attitudes, speeds and heights. The last three rows display more action samples of our dataset, showing the diversities, e.g., distinct views, various capture sites, weathers, scales, and motion blur.

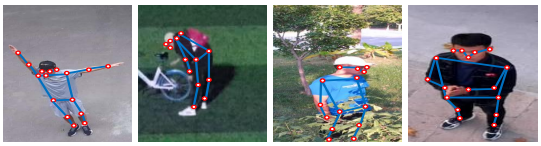


Figure 3: Samples for pose estimation.

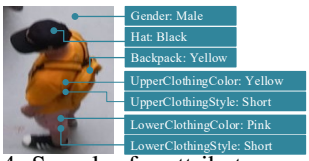


Figure 4: Samples for attribute recognition.

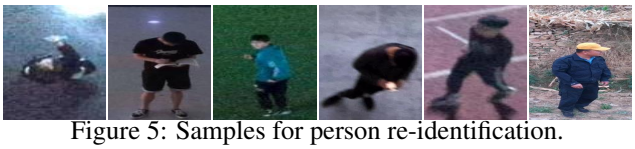


Figure 5: Samples for person re-identification.

(4) **ReID with various human poses and views.** Besides, we collect 41,290 image samples with 1,144 iden-

tities, using our UAV platform for person ReID. Captured from unique UAV perspectives, the person samples in this task also present very rich viewpoint, pose, and illumination variations, as shown in Figure 5.

We have obtained consent from all the captured human subjects to release all the tasks and samples in our dataset for non-commercial research and academic use.

3.3. Evaluation Criteria

Action Recognition. The cross-subject evaluation protocol is defined for action recognition. We use 89 subjects for training and 30 subjects for testing. The classification accuracy is used for performance evaluation.

Pose Estimation. We pick 16,288 frames from our manually-annotated frames for training and 6,355 frames for testing. Here mAP is used as the evaluation metric.

Person Re-identification. For person re-identification, we use 11,805 images with 619 identities for training, 28,435 images with 525 identities for gallery, and the rest 1,050 images are query images. To measure the re-identification performance, we use the mean average precision (mAP) and cumulative match characteristic top-k accuracy (CMC_k) as metrics.

Attribute Recognition. For attribute recognition, we use training and testing sets of 16,183 and 6,080 frames respectively. To evaluate the performance of attribute recognition, we measure the classification accuracy for each attribute.

4. Guided Transformer Network for Fisheye Video Action Recognition

Fisheye cameras are able to capture panoramic or hemispherical videos thanks to the ultra-wide view-angles [26, 3, 31]. Thus fisheye cameras are often equipped on UAVs to attain broad aerial views [10]. However, videos captured by fisheye cameras are often distorted, which makes fisheye-based action recognition challenging.

In this section, we propose an effective method, named Guided Transformer I3D (GT-I3D) to boost action recognition performance in fisheye videos as illustrated in Figure 6.

4.1. Guided Transformer I3D

As shown in Figure 6(a), the overall architecture of our method consists of two streams, namely a fisheye stream and a pre-trained RGB stream. We use I3D [4] as our baseline network considering its powerful representation learning capabilities. However, fisheye video samples are always distorted which brings difficulties for the original I3D to learn discriminative features.

Therefore, inspired by the spatial transformer network (STN) [12], we propose a guided transformer module, GT-Module, as shown in Figure 6(b). It can be inserted immediately before maxpooling layers in original I3D to warp each "pixel" on the feature maps extracted from distorted videos, by learning a series of unbounded transformations. The original STN [12] comprises a localization network learning transformation parameters ω from source features, a grid generator deriving transformations $\Phi(\omega)$ from the learned transformation parameters, and a sampler learning to map source features to target features based on the transformations.

Compared to original STN, we here use a 3D localization network replacing the original localization network, to learn 3D transformation parameters ω from fisheye video features f^F. Note that, to better alleviate the distortion problems, the learned parameters ω are unbounded, i.e. they are not normalized between -1 and 1 in our network. Then we use a grid generator to derive non-linear transformations $\Phi(\omega)$ conditioned on the learned parameters ω. Finally, a sampler is employed to shift each pixel (x_s, y_s) in the source fisheye feature maps f^F to (x_t, y_t) in the target feature maps f^T, as formulated in Eq. 1. Readers are also referred to the original paper of STN [12] for a detailed explanation of the mechanism.

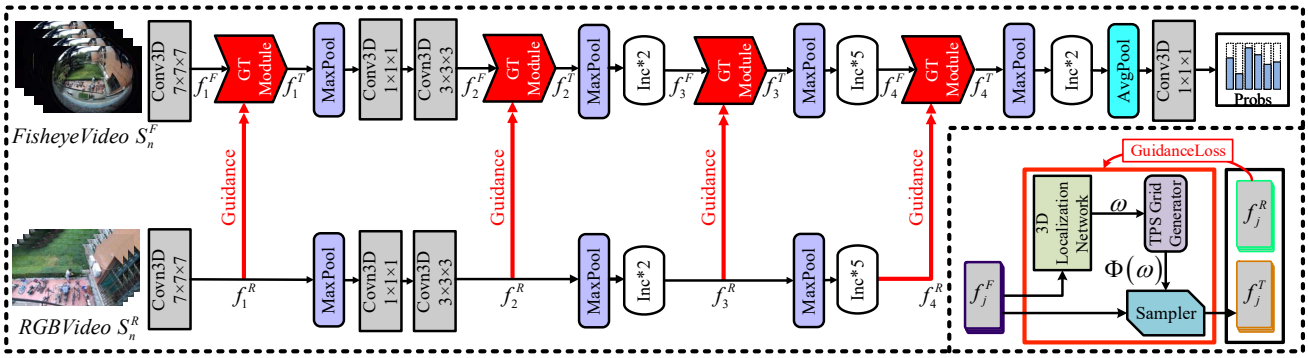
$$\begin{pmatrix} x_t \\ y_t \end{pmatrix} = \Phi(\omega) \begin{pmatrix} x_s \\ y_s \end{pmatrix} \tag{1}$$

However, it is still difficult to train the original STN to learn how to effectively handle the distortion problems, because there lacks explicit guidance for the warping feature. Thus, we propose to use RGB videos in our UAV-Human dataset as the guidance information to constrain the learning of the transformers by applying the Kullback–Leibler divergence loss between the flat RGB features f^R, and the transformed features f^T as shown in Figure 6.

It is worth noting that our GT-Module preserves the same shape among f^F, f^T and f^R with depth D, channel C, height H and width W, i.e. f^F, f^T and $f^R \in \mathbb{R}^{D\times C\times H\times W}$. In addition, the learned transformations vary among the D frames. This means if the input feature maps f^F contain D frames, the learned transformations also contain D frames accordingly, and within each frame, we follow the original STN [12] to apply the same transformations to preserve spatial consistency across all channels of this frame.

4.2. Training and Testing

Training. During training, the RGB and fisheye videos of the same action instance are respectively taken as inputs for the two streams. We uniformly sample n frames from each video as the input (S_n^R and S_n^F) for the RGB and fisheye streams respectively. Note that the RGB stream has been pre-trained on the flat RGB video samples in advance, and thus the RGB stream is fixed here to guide the training of the fisheye stream. As mentioned above, all the GT-Modules are inserted immediately before maxpooling layers of I3D. Therefore, the fisheye features f^F are passed



(a) GT-I3D framework (b) GT-Module details

Figure 6: (a) Illustration of our Guided Transformer I3D (GT-I3D) framework. The proposed GT-I3D consists of a fisheye stream that learns to rectify the distorted fisheye video S_n^F for action recognition, by using integrated Guided Transformer Modules (GT-Modules) and a pre-trained RGB stream (fixed) fed with flat RGB video S_n^R to guide the training of the GT-Modules. (b) Illustration of the detailed GT-Module. The 3D localization network is used to learn unbounded transformation parameters ω over the original fisheye features f^F. The grid generator is to derive the non-linear transformation $\Phi(\omega)$ from ω. The sampler is used to warp the distorted fisheye features f^F to the final transformed feature f^T based on $\Phi(\omega)$. The obtained f^T are further compared with the flat RGB features f^R to constrain the GT-Module to learn better transformations.

through each GT-Module to achieve transformed features f^T as shown in Figure 6. Meanwhile, the corresponding flat RGB features f^R are used to guide each transformer by applying KL divergence constraint between f^T and f^R as formulated below,

$$L_G = -\sum_{j=1}^{J} f_j^R \cdot (\log f_j^T - \log f_j^R) \tag{2}$$

where J is the number of inserted GT-Modules, i.e. the number of maxpooling layers in I3D, as shown in Figure 6. In addition, a standard classification loss is used to update the fisheye stream by comparing the predicted score $\hat{y}_k$ and the class label y_k as follows,

$$L_C = -\sum_{k=1}^{K} y_k \log \hat{y}_k \tag{3}$$

where K is the number of action classes. In summary, the following overall objective function can be used to update our GT-I3D framework.

$$L = L_G + L_C \tag{4}$$

Testing. During evaluation, only the fisheye video sequence S_n^F need to be sent to the fisheye stream to predict the action category, while the RGB stream is not required. Thus for testing scenarios, if the UAV is only deployed with a fisheye camera, we can still use this well-trained fisheye stream for more reliable action recognition without requiring flat RGB videos.

Implementation Details. To train our GT-I3D framework, the initial learning rate is set to 4e-3 and batch size is set to 16. For S_n^R and S_n^F, we choose $n = 64$ frames as the inputs. We use SGD as the optimizer to train our network on 2 Nvidia RTX2080Ti for 30K iterations.

5. Experiments

5.1. Evaluation on Action Recognition

Evaluation of multiple modalities. The proposed UAV-Human provides multiple data modalities for action recognition. Therefore we evaluate the performance of I3D network [4] on different data modalities and report action recognition results in Table 2.

We can observe that when using night-vision and IR videos as inputs, the highest accuracies of 28.72% and 26.56% are achieved. This is partially because a large portion of videos are collected in dark environments, while night-vision and IR videos attain clearer visual information in the nighttime compared to other modalities. Note that performance of depth sequences is slightly weaker than RGB videos mainly because depth sequences contain noises. Moreover, the performance discrepancy between fisheye videos and RGB videos is mainly caused by the intrinsic distortion issue of fisheye sensors.

Our proposed method is able to rectify distorted fisheye videos and thus improves the performance of original fisheye methods. The performance of our full model (23.24%) is even competitive to the RGB model (23.86%) as shown in Table 2. Moreover, ablation studies are also conducted. First we use guidance loss only, by applying KL divergence constraint directly on two feature maps encoded by the fisheye and RGB streams respectively, and attain performance of 21.68%. Then we use video transformers only, without guidance information from the RGB stream, and obtain performance of 21.49%. We can observe that our full model also achieves the highest accuracy among all fisheye-based methods. This indicates the efficacy of our proposed GT-Module.

Table 2: Evaluation of using different modalities for action recognition. All modalities achieve relatively low accuracies due to the practical challenges in our UAV-Human dataset, such as continuously changing of views, scales and locations and blurry images caused by different UAV flight attitudes, speeds and heights.

Modality	Accuracy (%)
RGB Video	23.86
Depth Video	22.11
IR Video	26.56
Night-vision Video	28.72
Fisheye Video	20.76
Fisheye Video+Guidance Loss	21.68
Fisheye Video+Video Transformer	21.49
Fisheye Video+Our Full Model	**23.24**

Table 3: The results of skeleton-based action recognition.

Method	Accuracy(%)
DGNN[29]	29.90
ST-GCN [38]	30.25
2S-AGCN [30]	34.84
HARD-Net [17]	36.97
Shift-GCN [6]	37.98

Table 4: The results of pose estimation.

Method	mAP(%)
HigherHRNet [5]	56.5
AlphaPose [9]	56.9

Note that, for RGB videos we also compare the performances of I3D baseline [4] and TSN baseline [37], and the recognition accuracies are 23.86% and 18.15% respectively.

Evaluation of the state-of-the-art methods on skeleton. Here we evaluate different state-of-the-art methods [38, 30, 29, 17, 6] on skeleton modality of the UAV-Human. As shown in Table 3, all skeleton-based methods outperform video-based methods shown in Table 2. The reason is that videos are sensitive to varying backgrounds, scales and locations led by the continuously changing of camera positions and viewpoints of the UAV. However, skeletal representations are more robust to complex backgrounds and additional normalization methods such as scaling and translating the skeletons [19] can be applied, which can make skeleton a more robust representation compared to other modalities in such a challenging UAV scenario.

5.2. Evaluation on Pose Estimation

Table 4 shows the results of two prevalent pose estimation methods [9, 5] on the UAV-Human dataset. We can observe that both methods attain relatively weak performance (56.9% [5] and 56.5% [9] respectively) on our UAV-Human dataset. This is possibly because that varying scales and views caused by multiple UAV attitudes plus diversified subjects' postures and complex occlusions bring more challenges to pose estimation in UAV application scenarios.

Table 5: Results of person re-identification.

Method	mAP	Rank-1	Rank-5
Part-Aligned [33]	60.86	60.86	81.71
PCB [34]	61.05	62.19	83.90
Tricks [20]	63.41	62.48	84.38
DG-Net [40]	61.97	65.81	85.71

Table 6: Results of Attribute Recognition. UCC/S and LCC/S represent Upper Clothing Color/Style and Lower Clothing Color/Style respectively.

Method	Accuracy (%)				
	Gender	Hat	Backpack	UCC/S	LCC/S
ResNet [23]	74.7	65.2	63.5	44.4/68.9	49.7/69.3
DenseNet [9]	75.0	67.2	63.9	49.8/73.0	54.6/68.9

5.3. Evaluation on Person Re-Identification

Three state-of-the-art person ReID methods [33, 34, 20, 40] are evaluated on our dataset. The results are presented in Table 5. Note our dataset is collected by moving cameras on a UAV, and the person images are often captured from overhead perspectives. This means our dataset brings brand new challenges to person re-identification, that can encourage the future deep neural networks to learn more representative features.

5.4. Evaluation on Attribute Recognition

We train two baseline methods using ResNet and DenseNet models as their respective feature extractors to identify common visual attributes. As shown in Table 6, recognition on clothing colors and styles achieve the lowest accuracies. This is possibly because that our dataset is captured in a relatively long period of time, and thus we have diversified subjects with different dressing types, plus large variations of viewpoints caused by multiple UAV attitudes, making our UAV-Human a challenging dataset for attribute recognition.

6. Conclusion

To the best of our knowledge, our dataset is the largest, most challenging and most comprehensive UAV-based dataset for human action, pose, and behavior understanding. We believe that the proposed UAV-Human will encourage the exploration and deployment of various data-intensive learning models for UAV-based human behavior understanding. We also propose a GT-I3D network for distorted fisheye video action recognition. The experimental results show the efficacy of our method.

Acknowledgements. This work is supported by SUTD Projects PIE-SGP-AI2020-02, SRG-ISTD-2020-153, the NSFC Grants U1913204 and 61991411, and the Natural Science Foundation of Shandong Province for Distinguished Young Scholars under Grant ZR2020JQ29.

References

[1] Ucf aerial action data set. https://www.crcv.ucf.edu-/data/UCF_Aerial_Action.php. 2

[2] Ucf-arg data set. https://www.crcv.ucf.edu/data/UCF-ARG.php. 2

[3] Mohammadamin Barekatain, Miquel Martí, Hsueh-Fu Shih, Samuel Murray, Kotaro Nakayama, Yutaka Matsuo, and Helmut Prendinger. Okutama-action: An aerial view video dataset for concurrent human action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–35, 2017. 1, 2, 6

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 6, 7, 8

[5] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5386–5395, 2020. 8

[6] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with shift graph convolutional network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 183–192, 2020. 8

[7] Konstantinos K Delibasis, Vassilis P Plagianakos, and Ilias Maglogiannis. Pose recognition in indoor environments using a fisheye camera and a parametric human model. In *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, volume 2, pages 470–477. IEEE, 2014. 3

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[9] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2334–2343, 2017. 3, 8

[10] Alex Gurtner, Duncan G Greer, Richard Glassock, Luis Mejias, Rodney A Walker, and Wageeh W Boles. Investigation of fish-eye lenses for small-uav aerial photography. *IEEE Transactions on Geoscience and Remote Sensing*, 47(3):709–721, 2009. 6

[11] Jian-Fang Hu, Wei-Shi Zheng, Jianhuang Lai, and Jianguo Zhang. Jointly learning heterogeneous features for rgb-d activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5344–5352, 2015. 3

[12] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015. 6

[13] Yanli Ji, Feixiang Xu, Yang Yang, Fumin Shen, Heng Tao Shen, and Wei-Shi Zheng. A large-scale varying-view rgb-d action dataset for arbitrary-view human action recognition. *arXiv preprint arXiv:1904.10681*, 2019. 3

[14] Vladimir V Kniaz, Vladimir A Knyaz, Jirí Hladuvka, Walter G Kropatsch, and Vladimir Mizginov. Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 3

[15] Quan Kong, Ziming Wu, Ziwei Deng, Martin Klinkigt, Bin Tong, and Tomokazu Murakami. Mmact: A large-scale dataset for cross modal human action understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8658–8667, 2019. 3

[16] Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *arXiv preprint arXiv:1806.11230*, 2018. 3

[17] Tianjiao Li, Jun Liu, Wei Zhang, and Lingyu Duan. Hard-net: Hardness-aware discrimination network for 3d early activity prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 8

[18] Chunhui Liu, Yueyu Hu, Yanghao Li, Sijie Song, and Jiaying Liu. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. *arXiv preprint arXiv:1703.07475*, 2017. 3

[19] Jun Liu, Amir Shahroudy, Mauricio Lisboa Perez, Gang Wang, Ling-Yu Duan, and Alex Kot Chichung. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 3, 8

[20] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 8

[21] Manuel Martin, Alina Roitberg, Monica Haurilet, Matthias Horne, Simon Reiß, Michael Voit, and Rainer Stiefelhagen. Drive&act: A multi-modal

dataset for fine-grained driver behavior recognition in autonomous vehicles. In *Proceedings of the IEEE international conference on computer vision*, pages 2801–2810, 2019. 3

[22] Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for uav tracking. In *European conference on computer vision*, pages 445–461. Springer, 2016. 1, 2

[23] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 8

[24] Francesco Nex and Fabio Remondino. Uav for 3d mapping applications: a review. *Applied geomatics*, 6(1):1–15, 2014. 1

[25] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, JK Aggarwal, Hyungtae Lee, Larry Davis, et al. A large-scale benchmark dataset for event recognition in surveillance video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3153–3160. IEEE, 2011. 2

[26] Asanka G Perera, Yee Wei Law, and Javaan Chahl. Uav-gesture: A dataset for uav control and gesture recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 1, 2, 5, 6

[27] Hossein Rahmani, Arif Mahmood, Du Huynh, and Ajmal Mian. Histogram of oriented principal components for cross-view action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 38(12):2430–2443, 2016. 3

[28] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016. 3

[29] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with directed graph neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7912–7921, 2019. 8

[30] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12026–12035, 2019. 8

[31] Amarjot Singh, Devendra Patil, and SN Omkar. Eye in the sky: Real-time drone surveillance system (dss) for violent individuals identification using scatternet hybrid deep learning network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1629–1637, 2018. 3, 6

[32] Veerachart Srisamosorn, Noriaki Kuwahara, Atsushi Yamashita, Taiki Ogata, Shouhei Shirafuji, and Jun Ota. Human position and head direction tracking in fisheye camera using randomized ferns and fisheye histograms of oriented gradients. *The Visual Computer*, pages 1–14, 2019. 3

[33] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 402–419, 2018. 8

[34] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 480–496, 2018. 8

[35] Chiara Torresan, Andrea Berton, Federico Carotenuto, Salvatore Filippo Di Gennaro, Beniamino Gioli, Alessandro Matese, Franco Miglietta, Carolina Vagnoli, Alessandro Zaldei, and Luke Wallace. Forestry applications of uavs in europe: A review. *International Journal of Remote Sensing*, 38(8-10):2427–2447, 2017. 1

[36] Lichen Wang, Bin Sun, Joseph Robinson, Taotao Jing, and Yun Fu. Ev-action: Electromyography-vision multi-modal action dataset. *arXiv preprint arXiv:1904.12602*, 2019. 3

[37] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2740–2755, 2018. 8

[38] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *arXiv preprint arXiv:1801.07455*, 2018. 8

[39] Shizhou Zhang, Qi Zhang, Yifei Yang, Xing Wei, Peng Wang, Bingliang Jiao, and Yanning Zhang. Person re-identification in aerial imagery. *IEEE Transactions on Multimedia*, 2020. 2

[40] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2138–2147, 2019. 8