Human Action Recognition in Drone Videos using a Few Aerial Training Examples

Waqas Sultani[1] and Mubarak Shah[2]

***Abstract*— Drones are enabling new forms of human actions surveillance due to their low cost and fast mobility. However, using deep neural networks for automatic aerial action recognition is difficult due to the need of humongous number of aerial human action videos needed for training. Collecting a large collection of human action aerial videos is costly, time-consuming and difficult. In this paper, we explore two alternative data sources to improve aerial action classification when only a few training aerial examples are available. As a first data source, we resort to video games. We collect plenty of ground and aerial videos pairs of human actions from video games. For the second data source, we generate discriminative fake aerial examples using conditional Wasserstein Generative Adversarial Networks. We integrate features from both game action videos and fake aerial examples with a few available aerial training examples using disjoint multitask learning. We validate the proposed approach on several aerial action datasets and demonstrate that aerial games and generated fake aerial examples can be extremely useful for an improved action recognition in real aerial videos when only a few aerial training examples are available.**

I. INTRODUCTION

Nowadays, drones are ubiquitous and actively being used in several applications such as sports, entertainment, agriculture, forest monitoring, military, and surveillance [1]. In video surveillance, drones can be much more useful than CCTV cameras due to their freedom of mobility and low cost. One critical task in video surveillance is monitoring human actions using drones.

Automatically recognizing human action in drone videos is a daunting task. It is challenging due to drone camera motion, small actor size and most importantly the difficulty of collecting large scale training aerial action videos. Computer vision researchers have tried to detect human action in varieties of videos including sports videos [2] , surveillance CCTV videos [3], cooking and ego-centric videos [4]. Furthermore, drones are recently being used to capture 3D human motion [5] and autonomously capturing cinematic shots of human action scenes [6]. However, despite being very useful and of practical importance, not much research work is done to automatically recognize human action in drone videos.

Deep learning models are data-hungry and need hundreds of training videos examples for robust training. However, collecting training dataset is quite challenging in several robotic vision applications such as semantic segmentation [7], measuring 6D object pose [8], and depth image classification [9]. Recently, computer graphics techniques and gaming technology have improved significantly. For example, GTA (Grand Theft Auto) and FIFA (Federation International Football Association) gaming engines use photo-realistic simulators to render real-world environment, texture, objects (human, bicycle, car, etc) and human actions. We propose to collect and use games action videos to improve human action recognition in real-world aerial videos. Games videos for action recognition are intriguing because 1) without much effort, one can collect a large number of videos containing environment and motion that looks very close to real-world, 2) It is easy [10] to get detailed annotations for action detection and segmentation which are otherwise very expensive to obtain, 3) Most of the gaming engines allow the players to capture the same action from the different views (aerial, ground, front, etc.,) at the same time. This means that we can easily collect multi-view dataset with exact frame-by-frame correspondence. All three advantages make gaming videos quite appealing for aerial action recognition where data collection is difficult and expensive.

Fig. 1: Figure shows examples of videos captured by UAVs. In each video, different human action is being performed. We aim to automatically recognize human action in these videos when only a few training aerial examples are available.

Another direction to address the scarcity of data is through generating fake video examples using generative adversarial network [11]. However, the quality of images and videos generated by GAN is not yet good enough to train deep networks [12]. Therefore, we propose to generate fake discriminative aerial features of different actions using conditional Wasserstein GAN. We believe that the fake aerial examples, when combined with a few real action examples, can help learn a more generalized and robust aerial action classifier.

In this paper, we propose to utilize game videos and fake

[1]Waqas Sultani is with the faculty of Computer Science, Information Technology University, Pakistan waqas.sultani@itu.edu.pk

[2]Mubarak Shah is with the faculty of Computer Science, University of Central Florida, USA shah@crcv.ucf.edu

generated examples to improve aerial action classification when a few real training examples are available. However, one of the key challenges is the disjoint nature of the problem. Video games are designed to address the interest of game playing audience and contain human motions and environments biased towards some few specific human actions. For examples, the majority of actions in FIFA games are related to playing a soccer game in a soccer field and the majority of actions in GTA are about fighting. Therefore, it is highly likely that classes of actions in games are different from the types of action classes we are interested to recognize in the real world. Similarly, it is not easy to generate good discriminative fake features for all types of action. However, our key idea is that despite different classes in games and real videos and the low-quality nature of fake aerial features, all three data types (games, real and fake) capture similar local motion patterns, human movements and human-object interactions, and, if integrated properly, can help learn more generalized aerial action classifiers. To achieve this, we combine games and fake examples with a few available real training examples using disjoint multitask learning.

Note that in this paper, we call the videos as ground action videos if the person making the videos is on the ground and the aerial videos are the one that are taken by UAVs. In summary, this paper makes the following contributions:

- We propose to tackle the new problem of drone-based human action recognition when only a few aerial training examples are available.
- We demonstrate the feasibility of game action videos for improving action recognition in real-world aerial videos.
- We show that game and fake action examples can help to learn a generalized action classifier through disjoint multitask learning framework.
- We present two new action datasets: 1) Aerial-Ground game dataset containing 700 human action video pairs (1400 videos), 2) Real aerial dataset containing actions corresponding to eight actions of UCF101.

II. Related Work

Human action recognition in videos is one of the most challenging and active vision problems [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [6], [5]. Classical approaches used hand-crafted features [13], [14] to train generalized human action recognition models that can perform well across different action datasets [15], [16].

With the resurgence of deep learning, several deep learning approaches have been proposed for action recognition. Simonyan et al. [20] proposed RGB and optical flow-based networks for action recognition videos. Both RGB and optical flow networks employ 2D convolution. Tran et al. [18] demonstrated the feasibility of 3D convolution for action recognition. In addition to presenting a new large scale action recognition dataset of 400 classes, Carreira et al. [17] proposed two-Stream inflated 3D ConvNet (I3D) that is based on 2D convnet inflation and demonstrated state of the art classification accuracy. Recently, an efficient action recognition framework is proposed by Chen et al. [21]. Furthermore, there has been an increased interest to train the generalized action recognition model using multi-task learning. Kataoka et al. [22] put forwarded a multi-task approach for the out-of-context action understanding. Similarly, Kim et al. [23] proposed disjoint multi-task learning to obtain improved video action classification and captioning in a joint framework.

Recently, Zhou et al. [5] proposed to analyze human motion using videos that are captured through a drone that orbits around the person. They demonstrated that, as compared to static cameras, videos captured by drone are more suitable for better motion reconstruction. Similarly, Huang et al. [6] presented a system that can detect cinematic human actions using 3D skeleton points employing a drone.

Although human action recognition is a quite active area of research in computer vision, there are not many research works dealing with aerial action recognition in the literature. Wu et al., [19] proposed to use low-rank optimization to separate objects and moving camera trajectories in aerial videos. UCF-ARG dataset [24] contain ground, rooftop and aerial triplets of 10 realistic human actions. This dataset is quite challenging as it contains severe camera motion and human in these videos occupy only a few pixels. Perera et al. [25] proposed to use human pose features to detect gestures in aerial videos. They introduced a dataset that is recorded by a slow and low-altitude (around 10ft) UAV. Although useful, their dataset only contains gestures related to UAV navigation and aircraft handling. Recently, Barekatain et al. [26] proposed a new video dataset for aerial view concurrent human action detection. It consists of 43 minute-long fully-annotated sequences with 12 action classes. They used single shot detection approach [27] to obtain human bounding boxes and then used the features within those bounding boxes for action classification.

Gathering large-scale dataset and its annotation is difficult, expensive and requires hundreds of human hours. To address this challenge, there is increasing interest in using synthetic data to train deep neural networks. Josifovski et al. [28] proposed to use annotated synthetic data to train instance-based object detector and object 3D pose estimator. Mercier et al. [8] used weakly labeled images and synthetic images to train deep network for object localization and 6D pose estimation in real-world settings. Carlucci et al. [9] proposed to use synthetic data for depth image classification. Recently, Richter et al., [10] designed a method to automatically gather ground truth data for semantic segmentation and [7] presented an approach to use those game annotations for semantic segmentation in real images. Finally, Mueller et al., [29] put forwarded photo-realistic simulators to render real-world environment and provide a benchmark for evaluating tracker performance.

In this paper, in contrast to the above mentioned methods, we address the problem of aerial action recognition when only a few training examples are available and propose a novel method to boost the classification accuracy using game and fake generated examples employing disjoint multi-task learning.

Fig. 2: Sample frames from our game action dataset. The first row shows the videos recorded from aerial camera while the second row depicts the same videos captured from ground camera.

III. Proposed Approach

In this section, we provide the details of our game action videos collection, the method to generate fake aerial examples and finally disjoint multitask approach where we train the aerial classifier using game aerial, fake aerial and a few real aerial videos in a unified framework.

A. Games Action Dataset

We employ GTA-5 (Grand Theft Auto) or FIFA (Federation International Football Association) for collecting a dataset. For each action, we record ground and aerial videos pairs i.e., the same action frames captured from both aerial and ground cameras. In total, we collect seven human actions including cycling, fighting, soccer kicking, running, walking, shooting and skydiving. Due to the availability of plenty of soccer kicking in FIFA games, we collect kicking from FIFA and the rests of the actions are collected from GTA-5. For each action, the dataset contains 200 videos (100 ground and 100 aerial) with a total of 1400 videos for seven actions. Note that most of the scenes and interactions in the video games are biased towards actions related to fighting, shooting, walking and running, etc. Therefore, the use of these games action videos for action recognition in real-world scenarios (which contain several different actions) is not straightforward.

In addition, due to the recent success of unsupervised image-to-image translations networks [30], we use pre-trained couple GAN from [30] to convert game videos into realistic-looking videos.

B. Fake Aerial Examples Generation

We generate fake aerial videos features employing Generative Adversarial Networks (GAN) [11]. Generative Adversarial Networks is a powerful tool to generate realistic-looking fake images and videos. However, the quality of fake videos is still far from being used to train a deep architecture for classification. However, several recent works demonstrated that GAN can be used to generate good discriminative features [12].

GAN consists of two networks: Generator and Discriminator. Generator tries to mimic the real data distribution and fools the discriminator by producing realistic looking videos or features while the discriminator job is to robustly classify real and generated (fake) video or features. Both Generator and discriminator can be simple multi-layer perceptrons. As compared to vanilla-GAN, in conditional GAN [31], both generator and discriminator are conditioned on auxiliary information. Auxiliary information can be video labels or some other video features. Our goal is to generate fake aerial visual features given the real ground features (auxiliary information). The objective function for our conditional GAN is given by:

$$\begin{aligned}\mathcal{L}_{cgan} = \mathbb{E}[logD(f_{r_a}|f_{r_g})] \\ + \mathbb{E}[log(1 - D(G(z, f_{r_g})|f_{r_g}))],\end{aligned} \tag{1}$$

where D represents discriminator and G represents generator, in $D(f_{r_a}|f_{r_g})$, f_{r_a} and f_{r_g} are real aerial and ground features respectively. These features are randomly sampled from given real aerial and ground features distributions. Note that we do not assume any correspondence between f_{r_a} and f_{r_g}. Given the noise vector z and f_{r_g} , generator tries to fool discriminator by producing fake examples.

To optimize the objective function, usually KL-divergence or JS-divergence is employed to reduce the difference between real and generated data distributions. However, one of the key limitations with KL-divergence or JS-divergence is that the gradient of divergence decreases with the increase of distance, and generator learns nothing through gradient descent. To address this limitation, Wasserstein GAN is introduced [32], which uses Wasserstein distance. WGAN learns better because it has a smoother gradient everywhere. Finally, to make Wasserstein distance tractable, the 1-Lipschitz constraint is used through gradient penalty loss [33]. The objective function of our conditional Wasserstein GAN (WCGAN-GP) is given by:

$$\begin{aligned}\mathcal{L}_{cwgan} = \mathbb{E}[D(G(z, f_{r_g})|f_{r_g})] - \mathbb{E}[logD(f_{r_a}|f_{r_g})] \\ + \mathbb{E}[(\left\|\nabla_m D(m, (G(z, f_{r_g}))\right\|_2 - 1)^2]\end{aligned} \tag{2}$$

where $m = tG(z, f_{r_g}) + (1-t)f_{r_g}$ and t is uniformly sampled between 0 and 1.

Our ultimate goal is to train discriminative action classifiers using fake features which are generated through generative networks. Although the above objective function generates realistically looking features, it does not guarantee

generating the discriminative features suitable for classification. To accomplish this, we first train soft-max classifiers using a few available real aerial examples. In order to enforce WCGAN-GP to produce discriminative features, we use classification loss computed over the fake aerial examples given as:

$$\mathcal{L}_{cl} = -E[logP(y_{r_g}|G(z, f_{r_g}); \theta)], \tag{3}$$

where $P(y_{r_g}|\mathrm{G(z},f_{r_g})$ denotes the probability of correct label prediction of generated examples. Since labels for real ground and fake aerial examples are the same, we use the labels of real ground (y_{r_g}) as ground truth.

Reconstruction loss Recent works such as [34] demonstrated that the quality of generated examples could be improved by using traditional reconstruction loss (L_2 or L_1) in addition to GAN loss. In our case, the reconstruction loss needs exact corresponding pairs of aerial and ground videos. Unfortunately, it is quite difficult to obtain such aerial-ground pairs in the real-world scenario. Our key idea here is that we can empower the generator with reconstruction loss that is computed over the game videos. This mean, in addition to fooling discriminator and producing discriminative examples, the generator needs to produce an exact aerial visual feature for the corresponding ground visual feature of game videos. As discussed in Section III-A, it is easy to collect plenty of such aerial-ground game video pairs. The reconstruction loss is given as:

$$\mathcal{L}_{rec} = \left\|(G(z, f_{k_g}) - f_{k_g}\right\|_1, \tag{4}$$

where f_{k_g} represent game ground features and $(G(z, f_{k_g})$ represents generated fake aerial game features.

The overall objective function for fake aerial examples generation is given by

$$\mathcal{L} = \mathcal{L}_{cwgan} + \mathcal{L}_{cl} + \beta\mathcal{L}_{rec}. \tag{5}$$

C. Aerial videos classification using Disjoint Multi-Task learning

Multitask learning improves the generalization capabilities of the model by effectively learning multiple related tasks. It has been used in several computer vision problems to learn the joint model such as; simultaneous object detection and segmentation [35], surface normal and pixel labels [36] and joint pose estimation and action recognition [37]. One of the limitations of multitask learning is the requirement of availability of multiple labels for each task for the *same* data. However, most of existing action datasets do not have such labels and hence restraining multitask learning on these datasets. To address this, recently disjoint multitask learning [23] is introduced. In the disjoint multi-instance framework, we can utilize different datasets to improve the generalization of the deep network.

In this paper, we propose to employ game aerial videos and fake aerial videos to perform disjoint multitask learning for action recognition. Since the two datasets are different (games and real) and secondly, we do not assume any common action classes, this fits well in context of disjoint multitask learning.

We have three tasks in total; action classification on real, fake and games videos. We first compute deep features of a few available real aerial videos and game videos using 3D convolutional neural network [21]. Furthermore, we obtain fake aerial features using the method described in Section III-B. We have two fully connected layers shared between all three tasks and one dedicated fully connected layer for each task.

Assume that the real, game and fake dataset videos features respectively are denoted by $r \in \mathcal{R}$, $k \in \mathcal{K}$ $m \in \mathcal{M}$ respectively . To denote ground or aerial videos, we use subscript a and g such that r_a and r_g represents the real aerial and ground video features respectively. We denote the number of actions classes in the real, game, and fake data as $\mathcal{N}$, $\mathcal{M}$, and $\mathcal{F}$ respectively. Note that we do not assume $\mathcal{M}$=$\mathcal{R}$=$\mathcal{F}$. We train all branches for the classification using softmax as a final activation function along with cross-entropy loss given by:

$$\mathcal{L} = -E[logP(y|f; \theta)], \tag{6}$$

where f denotes visual feature, $P(y|f)$ denotes the probability of true prediction of f and θ denotes the parameters to be learned

Our disjoint multitask framework is shown in Figure 3. Branch ①, ⑤, ⑨ are trained using ground truth labels of the real, game and fake data respectively. Branch ④ and ⑦ respectively predicts the real labels for the game and fake data; similarly branch ② and ⑧ predicts the games labels for real and fake data respectively; and finally branch ③, ⑥ respectively predicts the fake labels for real and game data.

We train different branches of multi-task framework using the aerial, game and fake data in iterations. First, using a few real aerial videos, we compute the loss for ①, ② and ③ only (see Figure 3), whereas ① predicts the real data labels, ② predicts the game data labels and ③ predicts the fake data labels. Note that the input to the network is a few real aerial videos only. Although we have ground truth labels for the real videos, we do not have the game and fake action labels for real videos due to disjoint nature of the problem. Therefore, we use the prediction of ⑤ (train for game ground truth labels) and ⑨ (train for fake ground truth labels) and consider them as ground truth for computing the classification loss of ② and ③. The loss function for this scheme is given by:

$$\begin{aligned}\min_{\Theta} \sum_{r_a \in \mathcal{R}_a} & \overbrace{\mathcal{L}(y_{r_a}, P(y_{r_a}|r_a))}^{①} \\ & + \overbrace{\mathcal{L}(\hat{y_{k_a}}, P(y_{k_a}|r_a)}^{②} + \overbrace{\mathcal{L}(\hat{y_{f_a}}, P(y_{f_a}|r_a)}^{③},\end{aligned} \tag{7}$$

y_{r_a} is ground truth labels of real aerial videos, $P(y_{r_a}|r_a)$ represents predicted real labels for real videos, $\hat{y_{k_a}}$ are the labels obtained from the layer trained with game ground truth labels (④)and $P(y_{k_a}|r_a)$ is predicted game action labels for

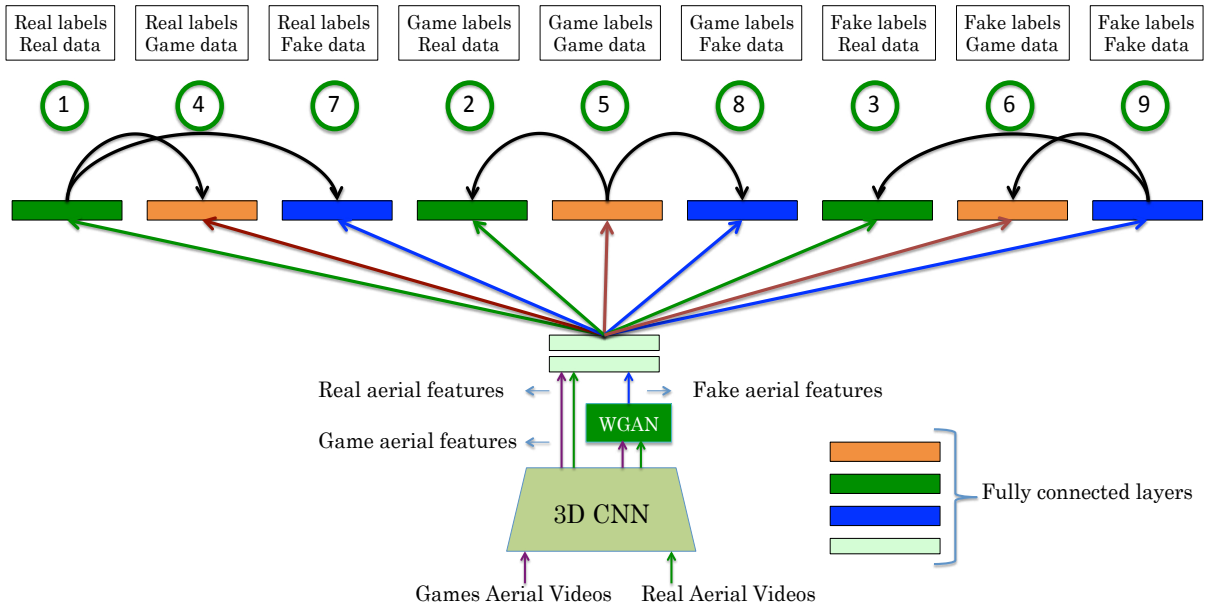

Fig. 3: Disjoint Multitasking Framework. Given the real and game and generated fake examples, we train different task specific layers in iterations. See the Section III-C for the detailed description of this figure.

real videos. Similarly, $y_{\hat{f}_a}$ are the labels obtained from the layer trained with ground truth labels for fake video (⑨), and $P(y_{f_a}|r_a)$ is predicted fake action labels for real videos and finally Θ represents network parameters.

After this, we train the networks using game aerial videos and compute the loss for ④, ⑤ and ⑥. ④ predicts the real data labels, ⑤ predicts the game data labels and ⑥ predicts the fake data labels. Note that the input to the network is game aerial videos and only ⑤ enjoys the ground truth game labels. As done before, we use the prediction of ① (train for real ground truth labels) and ⑨ (train for fake ground truth labels) and consider them as ground truth for computing the classification loss of ④ and ⑥. The loss function for this scheme is given by:

$$\min_{\Theta} \sum_{k_a \in \mathcal{K}_a} \overbrace{\mathcal{L}(y_{k_a}, P(y_{k_a}|k_a))}^{④} + \overbrace{\mathcal{L}(y_{\hat{r}_a}, P(y_{r_a}|k_a))}^{⑤} + \overbrace{\mathcal{L}(y_{\hat{f}_a}, P(y_{f_a}|k_a))}^{⑥}, \tag{8}$$

where y_{k_a} and $P(y_{a_k}|k_a)$ are ground truth and predicted action labels of game aerial videos, $y_{\hat{r}_a}$ is obtained from ① and $P(y_{r_a}|k_a)$ is predicted real action labels for game videos. Similarly, $y_{\hat{f}_a}$ is obtained from ⑨ and $P(y_{f_a}|k_a)$ is predicted action labels for fake data.

Finally, the loss function for the generated fake aerial data is given as:

$$\min_{\Theta} \sum_{f_a \in \mathcal{F}_a} \overbrace{\mathcal{L}(y_{f_a}, P(y_{f_a}|f_a))}^{⑦} + \overbrace{\mathcal{L}(y_{\hat{r}_a}, P(y_{r_a}|f_a))}^{⑧} + \overbrace{\mathcal{L}(y_{\hat{k}_a}, P(y_{k_a}|f_a))}^{⑨}, \tag{9}$$

where y_{f_a} and $P(y_{f_a}|f_a)$ are ground truth and predicted action labels of fake aerial videos, $y_{\hat{r}_a}$ is obtained from ① and $P(y_{r_a}|f_a)$ is predicted real action labels for fake videos. Similarly, $y_{\hat{k}_a}$ is obtained from ⑤ and $P(y_{k_a}|f_a)$ is predicted game action labels. Note that the number over the equations represents the corresponding numbers in Figure 3.

We repeat the above procedure for several epochs and fine-tune the parameters on the validation data. Note that we did not observe the forgetting effect [38] in our experiments.

IV. Experiments

The main goal of our experiments is to quantitatively evaluate the proposed approach and analyze the different components. To this end, we perform extensive experiments on the two aerial action datasets i.e., UCF-ARG [24] and YouTube-Aerial (collected by us).

A. Datasets

UCF-ARG[24]: UCF-ARG dataset contain 10 human actions. This dataset includes: boxing, carrying, clapping, digging, jogging, open-close trunk, running, throwing, walking and waving. This is multi-view dataset where videos are collected from an aerial camera mounted on Helium balloon, ground camera, and rooftop camera. All videos are of high resolution 1920 $\times$ 1080 and recorded at 60fps. The aerial

Method	Boxing	Carrying	Clapping	Digging	Jogging	Open-close-Trunk	Running	Throwing	Walking	Waving	Avg
Ground	33.3	00.0	20	00.0	40.0	00.0	60.0	6.70	20	0.00	18.0
DML using Games	53.3	13.3	00.0	26.7	26.7	41.7	66.7	6.70	00.0	13.3	24.8
DML using Games+Fake	13.3	26.7	00.0	00.0	6.70	58.3	66.7	6.70	40.0	80.0	29.8

TABLE I: Quantitative results for UCF-ARG dataset. Top row shows class-wise action recognition accuracy on aerial testing videos when training is done on video recorded by ground cameras. The second row demonstrates accuracy using disjoint multitask learning (DML) along with game videos. Finally, the last row demonstrates the same when game and fake examples are used together.

Method	Band-marching	Biking	Cliff Diving	Golf Swing	Riding Horse	Kayaking	Skateboarding	Surfing	Avg
Ground	33.3	26.7	93.3	53.3	6.00	53.3	26.7	93.3	48.2
DML using Games	53.3	73.3	66.7	100.0	13.3	66.7	53.3	86.7	64.2
DML using Games+Fake	66.7	33.3	86.7	80.0	60.0	66.7	40.0	86.7	65.0

TABLE II: Quantitative results on YouTube-Aerial dataset. Similar to the Table I, the top row shows class-wise action recognition accuracy on aerial testing videos when trained on ground videos, the second row demonstrates accuracy using game videos and third row shows the same when game and fake examples are used together.

videos contain severe camera shake and large camera motion. On average, each action contains 48 videos. We use 60% of videos of each action for training, 10% for validation and 30% for testing. Figure 4 shows some of the videos from UCF-ARG dataset.

YouTube-Aerial Dataset: We collect this new dataset ourselves from the drones videos available on YouTube, and will be made publicly available. This dataset contains actions corresponding to eight actions of UCF101 [2]. The actions include cycling, cliff-diving, golf-swing, horse-riding, kayaking, running, skateboarding, surfing, swimming, and walking. The videos in this dataset contain large and fast camera motion and aerial videos are captured at variable heights. A few examples of videos in this dataset are shown in Figure 1. Each action contains 50 videos. Similar to UCF-ARG dataset, we use 60%, 10%, and 30% of videos for training, validation, and testing respectively.

B. Implementation details

For visual features computation, we use 3D multi-fiber network [39]. Authors in [39] showed that multi-fiber network provides stat-of-the-art results on several competitive datasets and is the order of magnitude faster than several other video features networks. It achieves high computational efficiency by dividing the complex neural network into small lightweight networks. We extract the features (768D) for all videos from the second last layer of the network.

For disjoint multitask learning, we have two shared fully connected (f_c) layers (512 and 256 units respectively). We have nine task-specific layers: three f_c layers each with the number of units equal to the number of actions in the real dataset (shown in green color in Figure 3), games dataset (shown in orange color in Figure 3) and fake dataset (shown in blue color in Figure 3).

To generate fake examples, both our generator and discriminator contain four fully connected (f_c) layers where the first three f_c layers have Leaky ReLU activation. In the case of the generator, the last f_c has ReLU activation. The noise vector z (312D) is drawn from unit Gaussian. For all networks, we use Adam optimizer. The β in Equation 5 is chosen 0.001 and finally we weight the loss for fake data in Equation 9 with 0.1. For YouTube Aerial dataset, we weight the loss for game data in Equation 8 with 0.01. Note that all parameters are chosen over validation data. We use five videos of each action (named as a few available training examples in the above sections) in disjoint multitask learning. We ran the experiments several times with random initialization of the network (f_c layers) and report the average results.

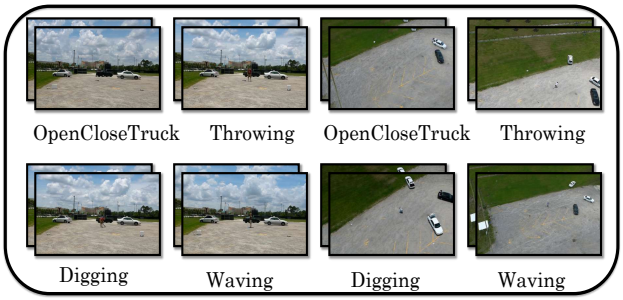

Fig. 4: Examples of videos from UCFAG dataset. The first two columns show the videos captured by the ground camera while the last two column show the same actions captured by a UAV.

C. Experiments Results

Table I and II demonstrate the experimental results on UCFARG and YouTube-Aerial datasets. All the classification results are of testing on real aerial videos. The first row demonstrates classification results when training is done on ground camera videos only. Note that UCFARG dataset contains ground cameras videos for the corresponding aerial actions videos. Some examples of ground camera videos are

shown in Figure 4. For YouTube-Aerial dataset, we use the videos of eight actions from UCF101 ground camera videos. The second row demonstrates the experimental results when the network is trained using games videos employing disjoint multi-task learning. Finally, the last row demonstrates the same when both games and fake data is used. In case of DML (Disjoint Multi-task Learning), we also use five available training examples as shown in Figure 3. As compared to YouTube-Aerial, low recognition accuracy on UCF-ARG dataset is due to videos containing non-discriminative backgrounds and very small actors size. It can be seen that for the most of the actions, the proposed approach results in improved action classification accuracy. The results emphasize the strength of the proposed approach and suggest that aerial video games and fake aerial features, when integrated properly, can improve the classification accuracy when a few training examples are available. Figure 5 shows the confusion matrix between different action for YouTube Aerial dataset. The results suggest that in addition to training on ground data only, using game and generated fake aerial examples significantly reduces the testing network confusion between different actions.

Comparison with Fine-tuning: We compute the classification accuracy when fine-tuning the network with only five aerial videos of each action without disjoint multitask learning. It can be seen in Table III that the proposed approach performs much better than simple fine-tuning the network. These results demonstrate the significance of disjoint multi-task learning using games and fake aerial data. Finally, the rightmost column demonstrates the upper bound classification accuracy when all training aerial examples (30 aerial examples for each action) are used.

	Few(5) Aerial	DML	All Aerial
UCF-ARG	17.3	29.8	32.5
YouTube-Aerial	60.8	65.0	68.3

TABLE III: Second column shows the classification accuracy when training is done only on five aerial videos. Third column shows proposed approach and the forth column depicts the accuracy when all training aerial examples are used (upper bound).

Impact of games on generating good fake aerial examples: To analyze the significance of using games for generating fake examples (Equation 4), we perform experiments with and without game reconstruction loss. The experimental results in Table IV demonstrate that the game reconstruction loss does help in generating good fake aerial examples. Note that to fully analyze the reconstruction loss contribution, in these experiments, we train the network with only fake aerial examples without employing disjoint multitasking learning.

	Without-Reconst	With-Reconst
YouTube-Aerial	61.92	63.33

TABLE IV: Classification accuracy on YouTube-Aerial dataset using fake aerial examples only. We observe the employing game reconstruction loss (right) produce more discriminative fake examples as compared to not using game reconstruction loss (left)

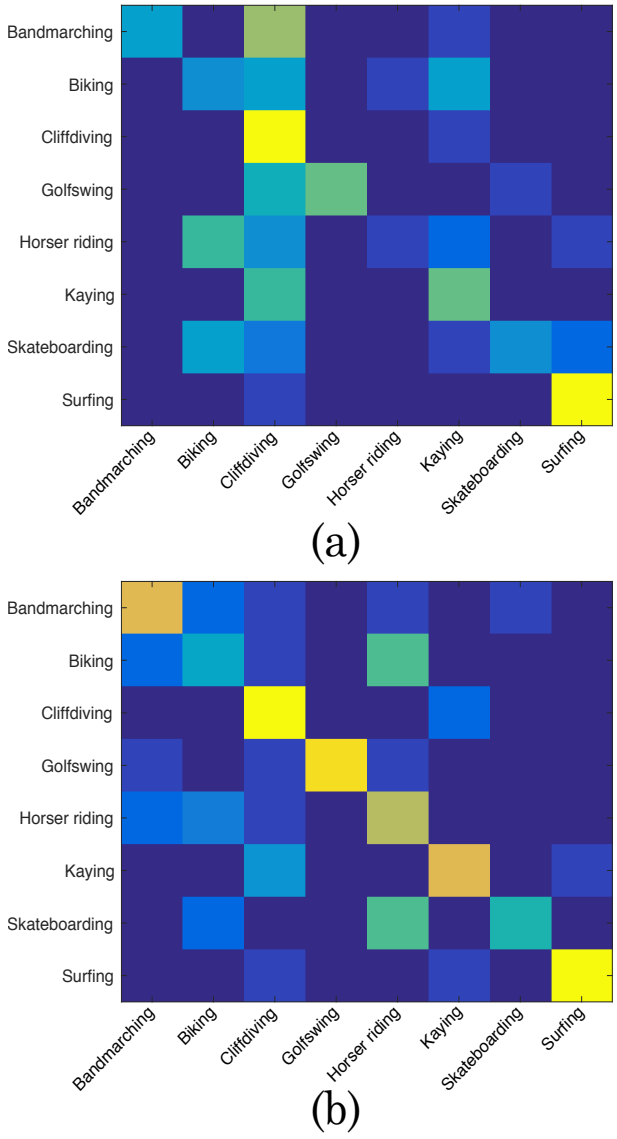

Fig. 5: Confusion matrix between actions in YouTube-Aerial dataset. (a) shows the confusion matrix when network is trained on ground videos. (b) shows the confusion matrix when network is trained using proposed approach.

V. Conclusion

Recently, low cost and lightweight hardware makes drones a good candidate for monitoring human actions. However, training the deep neural network for action recognition needs lots of training examples which are difficult to collect. In this paper, we explore two alternative data sources to increase the generalization capabilities of neural network classifiers. Our experimental results and thorough analysis demonstrated that game action videos and generated fake

examples, when integrated properly, can help to get improved aerial classification accuracy. The future works will aim at spatio-temporal localization of actor in drone videos, which will need granular deep features. Another future direction is to recognize human action in drone videos an unsupervised manner.

Acknowledgment: We like to thank Kaleem Khan and Tahir Khalil for collecting game action videos and YouTube-Aerial datasets. We also like to thank Sarfaraz Hussein and Mohsin Ali for the insightful discussions.

REFERENCES

[1] S. Dutta and C. Ekenna, "Air-to-ground surveillance using predictive pursuit," *2019 International Conference on Robotics and Automation (ICRA)*, pp. 8234–8240, 2019.

[2] K. Soomro, R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," in *ICCV*, 2013.

[3] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *2018 IEEE Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 6479–6488.

[4] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Scaling egocentric vision: The epic-kitchens dataset," in *European Conference on Computer Vision (ECCV)*, 2018.

[5] X. Zhou, S. Liu, G. Pavlakos, V. S. A. Kumar, and K. Daniilidis, "Human motion capture using a drone," *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2027–2033, 2018.

[6] C. Huang, F. Gao, J. Pan, Z. Yang, W. Qiu, P. Chen, X. Yang, S. Shen, and K.-T. Cheng, "Act: An autonomous drone cinematography system for action scenes," *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[7] W. Hong, Z. Wang, M. Yang, and J. Yuan, "Conditional generative adversarial network for structured domain adaptation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[8] J.-P. Mercier, C. Mitash, P. Giguère, and A. Boularias, "Learning object localization and 6d pose estimation from simulation and weakly labeled real images," *2019 International Conference on Robotics and Automation (ICRA)*, pp. 3500–3506, 2018.

[9] F. M. Carlucci, P. Russo, and B. Caputo, "A deep representation for depth images from synthetic data," *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1362–1369, 2016.

[10] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," *ArXiv*, vol. abs/1608.02192, 2016.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, 2014.

[12] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, "Feature generating networks for zero-shot learning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[13] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[14] H. Wang and C. Schmid, "Action recognition by dense trajectories," in *ICCV*, 2013.

[15] L. Cao, Z. Liu, and T. S. Huang, "Cross-dataset action detection," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010.

[16] W. Sultani and I. Saleemi, "Human action recognition across datasets by foreground-weighted histogram decomposition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[17] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[18] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, 2015.

[19] S. Wu, O. Oreifej, and M. Shah, "Action recognition in videos acquired by a moving camera using motion decomposition of lagrangian particle trajectories," in *Proceedings of the 2011 International Conference on Computer Vision*, ser. ICCV '11, 2011.

[20] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems 27*, 2014.

[21] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng, "Multi-fiber networks for video recognition," in *European Conference on Computer Vision (ECCV)*, 2018.

[22] H. Kataoka and Y. Satoh, "Unsupervised out-of-context action understanding," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019.

[23] D.-J. Kim, J. Choi, T. H. Oh, Y. Yoon, and I. Kweon, "Disjoint multi-task learning between heterogeneous human-centric tasks," in *Winter Conference on Application of Computer Vision*, 2018.

[24] "Ucf-arg data set," https://www.crcv.ucf.edu/data/UCF-ARG.php, accessed: 2020-09-8.

[25] A. Perera, Y. W. Law, and J. Chahl, "Uav-gesture: A dataset for uav control and gesture recognition," in *UAVision workshop, ECCV*, 2018.

[26] M. Barekatain, M. Martí, H. Shih, S. Murray, K. Nakayama, Y. Matsuo, and H. Prendinger, "Okutama-action: An aerial view video dataset for concurrent human action detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017.

[27] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," 2016. [Online]. Available: http://arxiv.org/abs/1512.02325

[28] J. Josifovski, M. Kerzel, C. Pregizer, L. Posniak, and S. Wermter, "Object detection and pose estimation based on convolutional neural networks trained with synthetic data," *2018 IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2018.

[29] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for uav tracking," in *Proc. of the European Conference on Computer Vision (ECCV)*, 2016.

[30] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 700–708. [Online]. Available: http://papers.nips.cc/paper/6672-unsupervised-image-to-image-translation-networks.pdf

[31] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *CoRR*, 2014. [Online]. Available: http://arxiv.org/abs/1411.1784

[32] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[33] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems 30*, 2017.

[34] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017.

[35] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., 2014.

[36] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert, "Cross-stitch networks for multi-task learning," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[37] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik, "R-cnns for pose estimation and action detection," 2014.

[38] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Dec 2018.

[39] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng, "Multi-fiber networks for video recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 352–367.