BOOSTING VIDEO-BASED PERSON RE-IDENTIFICATION WITH SYNTHETIC HUMAN AGENTS

SENTETİK İNSAN MODELLERİ İLE VİDEO TABANLI KİŞİ YENİDEN TESPİTİ BAŞARIMINI ARTTIRMAK

FIKRET KAYA

ASST. PROF. DR UFUK ÇELİKCAN

Supervisor

Submitted to Institute of Sciences of Hacettepe University

as a Partial Fulfillment to the Requirements

for the Award of the Degree of Master of Science

in Computer Engineering

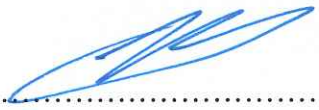
September 2019

This work named "**Boosting Video-Based Person Re-Identification with Synthetic Human Agents**" by **Fikret KAYA** has been approved as a thesis for the Degree of **Master of Science in Computer Engineering** by the Examining Committee Members mentioned below.

Prof. Dr. Haşmet GÜRÇAY

Head ..... 

Asst. Prof. Dr. Ufuk ÇELİKCAN

Supervisor ..... 

Prof. Dr. Pınar DUYGULU ŞAHİN

Member ..... 

Assoc. Prof. Dr. Mehmet Erkut ERDEM

Member ..... 

Asst. Prof. Dr. Elif SÜRER

Member ..... 

This thesis has been approved as a thesis for the **Degree of MASTER OF SCIENCE IN COMPUTER ENGINEERING** by Board Directors of the Institute of Graduate Studies in Science and Engineering on .....

Prof. Dr. Menemşe GÜMÜŞDERELİOĞLU

Direcor of the Institute of
Graduate School of Studies in Science

ETHICS

In this thesis study, prepared in accordance with the spelling rules of Institute of Graduate School of Science and Engineering of Hacettepe University,

I declare that

- All the information and documents have been obtained in the base of the academic rules
- All audio-visual and written information and results have been presented according to the rules of scientific ethics
- In case of using others works, related studies have been cited in accordance with the scientific standards
- All cited studies have been fully referenced
- I did not do any distortion in the data set
- And any part of this thesis has not been presented as another thesis study at this or any other university.

18 / 09 / 2019

FİKRET KAYA

YAYINLANMA FİKRİ MÜLKİYET HAKKLARI BEYANI

Enstitü tarafından onaylanan lisansüstü tezimin/raporumun tamamını veya herhangi bir kısmını, basılı (kağıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanması zorunlu metinlerin yazılı izin alarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

Yükseköğretim Kurulu tarafından yayınlanan "***Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge***" kapsamında tezim aşağıda belirtilen koşullar haricince YÖK Ulusal Tez Merkezi / H. Ü. Kütüphaneleri Açık Erişim Sisteminde erişime açılır.

- ☐ Enstitü / Fakülte yönetim kurulu kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren 2 yıl ertelenmiştir.
- ☐ Enstitü / Fakülte yönetim kurulu gerekçeli kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren .... ay ertelenmiştir.
- ☐ Tezim ile ilgili gizlilik kararı verilmiştir.

18 /09 /2019

FİKRET KAYA

ABSTRACT

BOOSTING VIDEO-BASED PERSON

RE-IDENTIFICATION WITH SYNTHETIC HUMAN AGENTS

Fikret KAYA

Master of Science, Computer Engineering Department

Supervisor: Asst. Prof. Dr. Ufuk ÇELİKCAN

September 2019, 42 pages

In recent years, research made in person re-identification has gained quite a bit of significance due to the increasing demand from a broad range of application fields with security and surveillance topping the list. A prominent part of this research utilizes deep learning methods that require large datasets with precisely extracted ground truth data. However, producing a large dataset from natural images for person re-identification poses many challenges. An alternative way of expanding the volume of available data is synthetically generating it. In this work, we present a synthetically generated dataset for video-based person re-identification that we created using real-world backgrounds and synthetically generated humanoids. Our dataset augments the DukeMTMC [12] dataset by simulating the scenes of the original dataset in our framework. Our dataset increases the size of the original dataset up to 3 times. This contribution improves the success rate of the Convolutional Neural Network based video based person re-identification approach by Wu et al. [34]. In addition to this, some tests conducted with the NVAN model of Liu et al. [23] to show that our method doesn't work in just one method, and we achieved similar achievements with this model as well. The results show that the improved dataset produced notably better results. Moreover, because of the generic format of our synthetic dataset generator framework, new datasets of different formats can be easily produced.

Keywords: Person Re-Identification, Synthetic Data Generation, Deep Learning, Automated Training Dataset Generation

ÖZET

SENTETİK İNSAN MODELLERİ İLE VİDEO TABANLI KİŞİ YENİDEN TESPİTİ BAŞARIMINI ARTTIRMAK

Fikret KAYA

Yüksek Lisans, Bilgisayar Mühendisliği

Danışman: Dr. Öğretim Üyesi Ufuk ÇELİKCAN

Eylül 2019, 42 sayfa

Son yıllarda özellikle güvenlik ve gözetleme gibi alanlardan gelen taleplerin artmasıyla birlikte kişi yeniden tespiti konusunda yapılan araştırmalar oldukça önem kazandı. Bu alanda yapılan araştırmaların önde gelenleri derin öğrenme tekniğini kullanıyor. Derin öğrenme teknikleri büyük veri setlerine ve bu veri setlerinin hassas bir şekilde çıkarılmış referans verilerine ihtiyaç duyar. Fakat kişi yeniden tespiti için kullanılacak bu veri setlerini doğal ortamlardan elde etmek bir takım zorluklar ortaya koyuyor. Bu zorlukları aşmanın bir yolu ise veriyi sentetik olarak üretmek veya olan veriyi sentetik veri ile çoğaltmak. Bu çalışmada, video tabanlı kişi yeniden tespitinde kullanılmak üzere gerçek sahne arka planları ve sentetik olarak üretilmiş insansı modelleri kullanarak ürettiğimiz veri setimizi sunuyoruz. Yarattığımız sistem ile DukeMTMC [12] veri setindeki sahneleri sanal ortamda yeniden oluşturup bu veri setini taklit eden bir sentetik veri seti oluşturuldu. Kurduğumuz sistem ile gerçek veri setinin 3 katı büyüklüğünde bir sentetik veri seti oluşturmayı başardık. Bu veri seti ile Dönüşümlü Sinir Ağı tabanlı bir kişi yeniden tespiti yönteminin (Wu et al. [2]) başarı oranında önemli oranda artış elde ettik. Oluşturduğumuz sentetik veri seti üretme sisteminin genel yapısı sayesinde farklı formatlarda ve ortamlarda hazırlanmış veri setlerinin simüle edilmesi de oldukça kolay.

Anahtar Kelimeler: Kişi Yeniden Tespiti, Sentetik Veri Seti Üretimi, Derin Öğrenme, Otomatik Eğitim Veri Seti Üretimi

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my supervisor Asst. Prof. Dr. Ufuk Çelikcan forhisprecious advice and leading. He supported me from the beginning of the thesis to the final stage with their valuable knowledge and experiences. Besides, I would also like to thank Assoc. Prof. Dr. Erkut ERDEM and Assoc. Prof. Dr. İbrahim Aykut Erdem for their guidance and support.

In addition to my supervisor, I would like to thank my thesis committee members, Prof. Dr. Haşmet GÜRÇAY, Prof. Dr. Pınar DUYGULU ŞAHİN, Asst. Prof. Dr. Erkut ERDEM and Asst.Prof. Dr. Elif SÜRER for reviewing this thesis and sharing their supportive feedback.

Finally, I would like to thank my family for their endless support, faith and contribution they have shown me throughout my education life. Every time I stumbled or had difficulty, they were there for me and motivated me to move on. In addition, the support of my friends during my studies cannot be underestimated, I am also grateful to them.

This study is part of the project “Using Synthetic Data for Deep Person Re-Identification” supported by TUBITAK 1001 program (Project No. 217E029).

CONTENTS

FIGURES

TABLES

1. INTRODUCTION

Person re-identification aims at spotting the person-of-interest from different cameras [34]. Thus, it is indispensable for security and surveillance systems. Therefore, both new algorithms and new datasets are being proposed to solve this problem. However, there are some issues that make this problem extremely difficult such as inter-camera viewpoint differences, illumination changes, pose variations [12] etc. The algorithm cannot handle these completely on its own. In this case, a dataset containing these features is needed. In order to have a good dataset not only these features, but also a sufficient number and variety of data are required. However, it is not very compelling to generate a dataset with these requirements with manpower. The general approach is to use a person detect or to retrieve humans' bounding boxes from every frame and all view-points and then label these humans. Since the size of the dataset is important in terms of both number of identities and bounding boxes [13], it is a very labor intensive job for man.

This job is very time consuming for men but it is effortless fora machine. The new trend to generate labeled datasets is based on computer graphics. The scene of the desired dataset can be created in a digital environment. In this digital scene we can easily use synthetically generated objects to create a highly detailed and precisely labeled dataset. The most important benefit of the dataset being digital is that the ground truth of the dataset is known in advance which allow us to achieve reputable results. This approach is widely used by the recent research and the results obtained prove this to be a most effective approach [8], [7], [14], [9]. This approach is being used especially in Convolutional Neural Networks because they bring thoroughly successful results with a large amount of good training data and the mentioned synthetic dataset generation technique is excellent at providing it.

In this work we used DukeMTMC-VideoReID [34] which is a subset of the DukeMTMC tracking dataset [28] for video-based person re-identification. This subset dataset contains 702 identities for training and another 702 for testing. The DukeMTMC dataset consists of 8 different cameras. Most of the identities appear in more than one camera. The identities used in training phase appear in2196 different videos, so on

average each identity exists in 3 different camera viewpoints. This size is remarkable because this dataset was created manually, spending quite a long time.

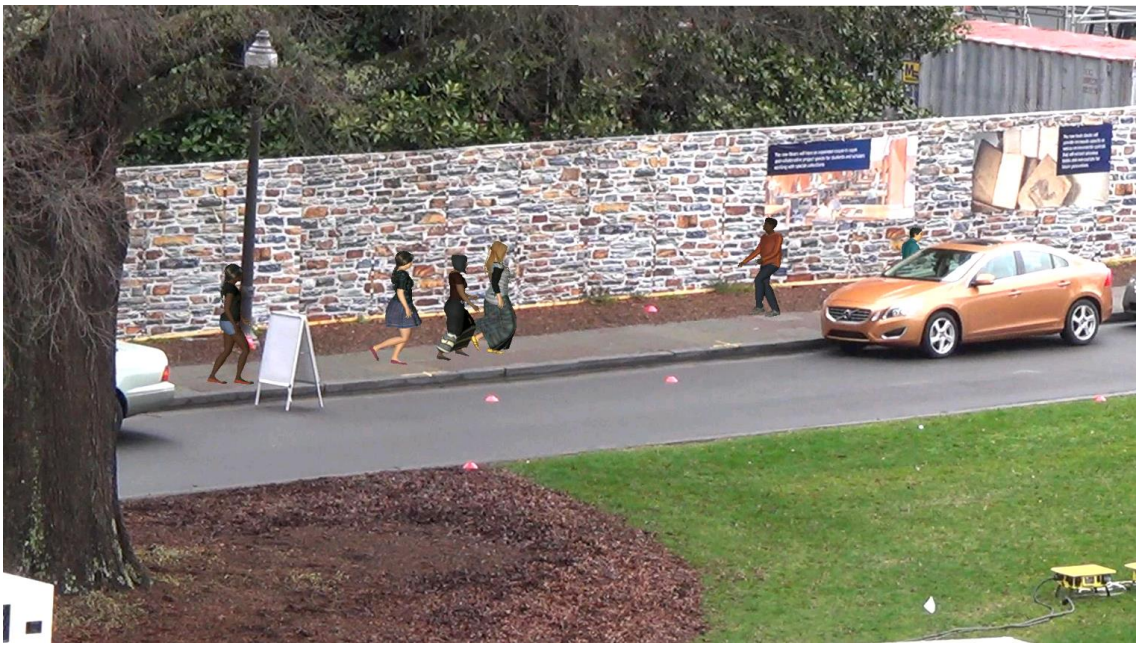


Figure 1.1: A sample image of our augmented data in 1920x1080 size. The scene background is obtained from Duke Dataset's scene 1. The foreground objects are our 6 synthetically generated humanoids (4 female, 2 male).

In this paper we present a new large scale synthetically generated dataset based on DukeMTMC-VideoReID [34] dataset. Our synthetic dataset consists of the same background as the real dataset. We obtained empty scenes that do not contain any human from the 8cameras. Using MakeHuman software [2], virtual human agents are produced in a wide range. A total of 1710 different characters are created. These humanoids were subjected to various operations and used on the scenes. Unity Game Engine [3] is used to perform these operations and generate the dataset. Then, a post processing operation is carried out and the synthetic dataset is successfully finalized.

The success of the generated synthetic dataset is tested with the convolutional neural network method mentioned in Wu's paper [34]. They propose 2 approaches both based on ETAP-Net model, one is supervised and the other one is semi-supervised method which utilizes a stepwise learning method. This approach focuses on video-based

person re-identification. They propose a method to improve the discriminative capability of the Convolutional Neural Network (CNN) feature representation by stepwise learning. The detailed description of this method will be provided in the results section. Results show that our photo realistic synthesized dataset has a huge impact on the results and increases the success rate of the CNN used for person re-identification method considerably. Apart from ETAP-Net model, we conducted some tests with Liu et al. NVAN model [23] as well. They target at refining the intermediate features as well as high-level features with non-local attention operations for this Non-local Video Attention Network (NVAN). They use this network to incorporate video characteristics into the representation at multiple feature levels. Thus, we utilized all these methods and our synthetic dataset and conducted experiments to show our positive contributions to the accuracy. Our detailed experiments conducted with ETAP-Net's supervised method; however to show that the proposed dataset's success, the same experiments conducted with ETAP-Net semi-supervised method as well. And then, in order to prove that our success is not limited to ETAP-Net model, we also did some test with NVAN model.

Figure 1.2: DukeMTMC-VideoReID Dataset contains 702 different people. These people appear in some of the 8 different scenes in Duke University.

In the following sections we will summarize some related works done in this domain. Then, the technical details of our study will be included with supporting samples and figures. Next, we will explain the detailed results we have achieved in person re-identification using our synthesized dataset, and finally give some advice and share our ideas for future work.

2. RELATED WORKS

2.1. MixedPeds Person Detection Dataset

Synthetic dataset generation approach has a widespread use and can be applied to solve various problems. Cheung [8] used this approach for pedestrian detection in unannotated videos.

Figure 2.1: An Example image from MixedPeds Dataset [8]. Synthetically generated static human-agents placed on real world background images.

They are using real-world background images and virtual agents to create the dataset. Their approach is designed to make as few assumptions as possible. They compute two main needs for their approach automatically from the unannotated dataset. Using a detector trained with KITTI and CALTECH dataset they produce a set of confident bounding box which is used to estimate the scale of pedestrians. Since the height of the bounding box changes with respect to the distance with camera, they compute scale ratio. The scale ratio is computed with detected bounding box height and a vanishing point method [10]. This scale ratio and vanishing point pair is used to calculate synthetic pedestrians' sizes that will be overlayed in a varied distance from the camera. Another information generated from detected bounding boxes is Spawn Probability Map, which is utilized to place the synthetic pedestrians at appropriate coordinates.

Using detected bounding boxes a 2D histogram which holds the probability ratio of the pixels is constructed. They also apply a Gaussian kernel to the histogram to increase the probability of the neighboring pixels. Then, these probabilities are used to spawn synthetic agents in image. In this way, they avoid placing synthetic agents on undesired places such as on top of an obstacle.

MixedPeds dataset contains 7 different virtual humans along with various number of clothes and skin colors. These humans are combined with clothes and skin colors and 9 different poses are generated for each of them. Synthetic humans are generated and mixed with a small set of real data. They indicate that using MixedPeds dataset as training data, they achieved up to 21.9% increase in accuracy. Which clearly shows how effective this approach is and can be used to increase learning rate and decrease the difficulty of finding a proper dataset which is suited to the problem definition.

2.2. Scene Specific Pedestrian Detectors without Real Data

Sometimes surveillance systems may require a detector that will be installed in a specific location. In such circumstances, this specific location request may make it difficult to find a real dataset suitable for the location for the relevant problem. As a solution to this drawback, Hattori et al. decided to use a set of customizable virtual pedestrians [14].

There are 2 main factors affecting the success of the dataset. The scene similarity of the training and test phases is an important factor affecting the success of the detector being developed. Another factor affecting the success rate is the camera calibration information provided by the dataset. The calibration data is an important asset for the detector. If camera calibration used in training and testing sets do not match, the success rate is negatively affected. Therefore in the optimal condition we should have the learning and testing data with the same background and camera calibration. However, this is not very suitable for real life application. We cannot gather real data to train detectors for each location to be used. We cannot do the same every time a camera or a

lens is changed as well. However, using camera calibration and scene geometry it is possible to generate a synthetic training dataset.

Figure 2.2: Synthetic pedestrians placed on grid locations

Using scene information such as the geometrical layout of the scene, locations where pedestrians may appear and obstacle positions the virtual dataset can be generated. This information is basically used to distribute the synthetic pedestrians to the scene correctly. According to the layout of the scene, pedestrians' geometrically correct rendering is generated. In addition to this, occlusions in the scene are also taken into consideration. Obstacle objects and walls in the scene where models may be occluded partially or completely are marked and occlusion is simulated during generation when it is needed.

A total of 36 virtual human models are used and these models are simulated in 3 different walking configurations in every location available in the scene. In every location, pedestrians are oriented 12times. As a result, about several million training images are obtained. The performance of this dataset is evaluated on Town center [4],PETS 2006 [30], and CMUSRD [15] datasets using a number of pedestrian detectors. The results show that for both 0.5 and 0.7 overlap criteria, the method of using purely synthetic human data outperforms all baselines with an AP of 0.90. The

second and third best performances are 0.86 and 0.73. These results show that a purely synthetic dataset can be used to train a person detector quite successfully.

2.3. Learning from Synthetic Humans

Estimating human pose or action requires a large amount of sensitively labeled training data. Recent advances in human pose estimation uses convolutional neural network methods which need extensive datasets. Finding the desired dataset is quite challenging, and creating it is even more challenging. Creating such a dataset manually is time consuming and it is difficult to extend it. Therefore, Varol et al present SURREAL (Synthetic Humans for Real Tasks); anew large scale dataset with synthetically generated humans rendered from 3D sequences of human motion capture data [31].

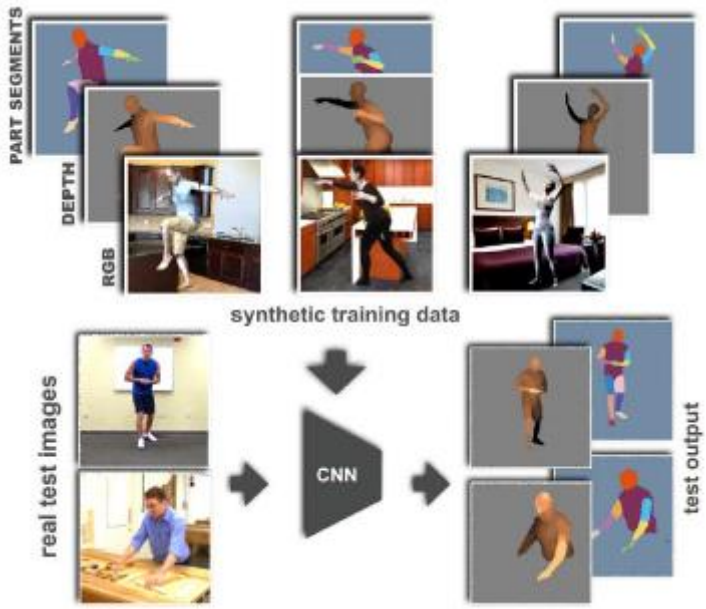


Figure 2.3: Synthetic models to train pose estimation algoritm.

Synthetic human models are created with SMPL body model [24].Then in order to render realistic body shapes, CAESAR dataset **issued** [29]. CMU MoCap database is utilized to generate realistic poses [1]. Then, the models are textured with CAESAR textures. Asa result, 115 subjects for training and 30 for testing are created. The dataset

contains 67.582 clips with 6.536.752 frames. Automatically generated ground truth consists of model's pose, depth maps, optical flows and segmentation masks.

The SURREAL dataset is tested and compared with several real datasets. In addition, in some cases the SURREAL dataset was tested by combining it with some real datasets as well. Their results show that although the synthetic dataset outperforms the real datasets the best performance is obtained by fine tuning the synthetic dataset with some real images.

3. METHODOLOGY

In this section of the paper, we will describe our approach to generate a synthetic dataset to train a convolutional neural net-work for person re-identification. We first describe the DukeMTMC-VideoReID dataset [34] which is a subset of DUKE MTMC dataset [28] because our dataset is derived from this dataset. Then implementation stages of our synthetic dataset generator will be explained in detail. We divided this into three parts; synthetic human generation stage, animating generated humans on proper positions in the scenes, and the post processing stage. In the post processing stage the final editing is made and the ground truths are created.

3.1. DukeMTMC-VideoReID Dataset

The DukeMTMC-VideoReID dataset is a special dataset derived from the Duke MTMC dataset to be used to train video based person re-identification applications.

The DUKE MTMC stands for multi target multi camera and DUKE stands for Duke University. This dataset was created in the Duke University Campus. Therefore the main area of utilization of the DUKE MTMC is multi target tracking in multi camera networks. The dataset contains more than 2 million frames which contain more than 2700 identities. The designated area on the campus was equipped with 8 static cameras and each camera recorded the area about 85 minutes in full hd at 60 fps. The ground truth of the dataset is manually extracted by 5 people over a year. This ground truth provides the following information for every frame of every identity:

- Human id,
- Frame index in the current camera and camera index,
- Bounding box boundaries,
- World coordinates

Person re-identification has drawn increasing attention in the last decade and as a result of this, researches made in this subject have grown considerably. To support these applications, Gou et al. created a new person re-identification dataset, DukeMTMC-

VideoReID by using Duke MTMC dataset [12]. Therefore we can consider DukeMTMC-VideoReID as a subset of the Duke MTMC dataset. To generate DukeMTMC-VideoReID dataset the bounding boxes provided by the superior dataset is used and bounding boxes of each person cropped out from the entire frame. To simulate real world systems, bounding box cropping process is made according to the off-the-shelf person detectors proposed by Market1501 [36] and CUHK03 [21] datasets. As a result, 1852 unique identities are included in the new dataset. 1413 of 1852 people appears in more than 3 camera's viewpoint. The dataset is divided into 3 parts; training, testing, and distractors. The 702 of 1852 identities are split as training data, another 702 as testing and remaining 48 are the dis-tractors. The people chosen as training data appears in 2196 videos so approximately each person appears in 3 different scenes. In addition to that, these 702 training identities consist of 369.656 frames in total.

Figure 3.1: Sample scenes from Duke MTMC Dataset. As seen clearly, the camera positions, angles and coverage areas in the scenes are different from each other.

3.2. Photo Realistic Synthetic Human Generation

The most commonly used 3D human model generator is the MakeHuman Software [2] which is an open source project. MakeHuman makes it possible to synthesize different realistic 3D human models and it also provides the joints' positions as well [27]. In this way we can use these models in different poses. MakeHuman provides many parameters to diversify the human models. We can encounter these virtual human models in many different ways, for example in games, simulations, animations, training etc. Recently we have encounters with these virtual humans in deep learning research, they play an active role as training data in these projects. For instance, Rahmani et al. used MakeHuman to synthesize human models and they fitted these virtual humans to real motion capture data and created their dataset to train their Non-Linear Knowledge Transfer Model (R-NKTM) for human action recognition [27].

Figure 3.2: Sample scenes from DukeMTMC4ReID Dataset. Every grid is an identity frame from dataset. The First row shows 5 sequential frames of a woman from scene 1 and the second row shows 5 sequential frames of a man from scene 2

We utilized MakeHuman software to synthesize the human models we need to train a person re-identification model. Therefore, besides being realistic, the size and variety of our dataset is very important for us. In order to diversify the human models, we used the following main parameters; age, gender, muscle, weight, height, proportion and ethnicity [6]. In addition to these, MakeHuman also provides some assets to enrich the data created such as; clothes, hair, facial hair, skin and eye materials, rig presets etc. Using the main parameters and assets we created 560 synthetic humans. We exported these humans in Filmbox (.fbx) format. Which is a format supported by the Unity Game Engine.

Besides the main parameters of MakeHuman, the contribution of the additional assets made great impact on the diversification of the humans. These assets are listed in Table 3.1 in detail.

Asset List		
Assets and Amounts	Male	Female
Hair and Facial Hair Styles	5 + 2	14
Bottom-Top Suits	8	6
Bottom Clothes	9	18
Top Clothes	8	18
Shoes	6	7

Table 3.1: Number of assets used to generate synthetic male and female models. Various combinations with these assets have been made and dressed in models.

In the synthetic human generation phase, a python script has been written to generate human models. First of all the ids of these assets are listed and to create noticeable differences between the models, these assets have been used with various combinations via their ids. In addition to these, the main parameters of the MakeHuman, age, gender, muscle, weight, height, proportion, ethnicity, skin color, are used to synthesize distinctive humans. Then these wearable combinations and randomized parameters are

put together and human model files are generated in MakeHuman format (.mhm). MakeHuman loads 560 .mhm files and converts them from plaintexts to 3D human models in Filmbox (.fbx) format. These exported human models are ready for Unity Game Engine and no additional process is needed to be imported by Unity.

Figure 3.3: DukeMTMC-VideoReID's real identities vs synthetically generated humanoids.

3.3. Animating Synthesized Humans on Scenes

In order to create a synthetic dataset, utilizing a game engine is not a new idea, it is a proven and successful method [17], [19], [25].Therefore we decided to use Unity 3D Game Engine [3] to render our models and create the core data of our synthetic data.

The process of animating synthesized human models in front of the scenes is a multi-phase process. This entire process is implemented in Unity. It begins with importing the human models to Unity. First of all, the animation types of models are set to humanoid. Then rigid body components are added to the humanoids to put their motions under the control of Unity's physics engine [3]. Then, an animation controller is created with 5 different walking animations, humanoids get these walking animation types one by one and in this way walking diversity is established. Finally, our main script animates the

humanoid in front of a white board with the same size of the real dataset, 1920x1080. While the humanoid is walking in front of this white scene, the screen shots of this process are taken with some information.

Animating the humanoids in front of the white scene is a rather complex operation in several aspects. First of all, selecting the right position is complex because there might be some obstacles. The second problem is the scenes are in 3D but if we animate the humanoids in 2D the realism decreases. Another problem is that occlusion objects in the scene; there are trees, cars, walls and sign boards that people walk behind so they are occluded, the same effect must be applied to the synthetic data as well.

Figure 3.4: These 9 models have been synthesized using MakeHuman. As seen here, combining clothes with each other and randomizing the main parameters of the MakeHuman is very useful to differentiate the models.

Figure 3.5: Concatenated images of a humanoid animated (walked) in front of the fifth scene.

In order to overcome the first problem, selecting the paths forhumanoids, we make use of the real dataset's ground truth. TheDukeMTMC dataset provides the 2D position of the people walkingin the scenes. Therefore, for every time a humanoid needs to walk,we selected a random path from the real dataset and animated thehumanoid in x-y plane accordingly. Using this data we also decidedto how to rotate our humanoids while walking. Calculating theangle between point x-y in the current frame and next 60th framegave us how to rotate the humanoid in y axis (yaw axis).

Although rotation in the y axis is calculated, the realism in thescene not established yet because cameras are not exactly parallelto the scene. To solve this situation we manually find the normals ofthe 8 scenes. Then we set x (pitch axis) and z (roll axis) axes of thesenormals to the humanoids in the initialization phase. Since we donot change these x and z angles (in world space) during movement,the animation looks very realistic. Only the local y angle is changedto rotate humanoid to its target point during walking in the scene.

Scenes are not fictional spaces but they are real campus scenes ofDuke University campus. Thus, there are a few occlusion objectsin the scenes. To include this effect in

our synthetic dataset wecome up with this idea; we pointed occlusion objects bottom centerpoints and measure their widths and heights to check whether thehumanoid is inside any of the occlusion areas. If so, the name of theocclusion is added to the ground truth data of that frame. In the postprocessing phase, using this information we are solving this occlu-sion problem. An example of occlusion included in frame's groundtruth is "camera1-7141-00066-signboard1"; "camera1" indicates thescene no is 1, the "7141" indicates the humanoid id, "00066" indicatesthe frame index of the current video, and remaining "signboard1"indicates the occlusion. In this frame the signboard must occludethe humanoid. This needs to be done in the post processing. As aresult of this phase, we get images containing animated humanoidsin front of white backgrounds and ground truth information.

3.4. Post Processing

In the post processing phase the main idea is to convert the images created in Unity to the DukeMTMC-VideoReID [34] format. TheDukeMTMC-VideoReID dataset is split to 3 folders; train, query and gallery. In the query set, there are 702 videos, extracted as frames. In the training, there are another 2196 videos, extracted as frames, for different identities. On average every identity in training is contained in 3 videos. These videos are not full size (1920x1080) images, they are extracted by the bounding boxes of the identities. Therefore the image sizes in this dataset are not the same for all.

The phase we call post processing is the core part of our frame-work. In this part we get the images created in Unity, the ground truth data and the occlusion overlay images. The occlusion overlay images are the occlusion objects we manually extracted from the real dataset's scenes. There are 15 occlusion objects in total. Using an image editor, the occlusion objects are selected manually and they are extracted from the background images, so the new images only include the occlusion objects, besides that they are transparent. These transparent background occlusion images are overlaid on top of the humanoid images, in this way partial or full occlusion is established.

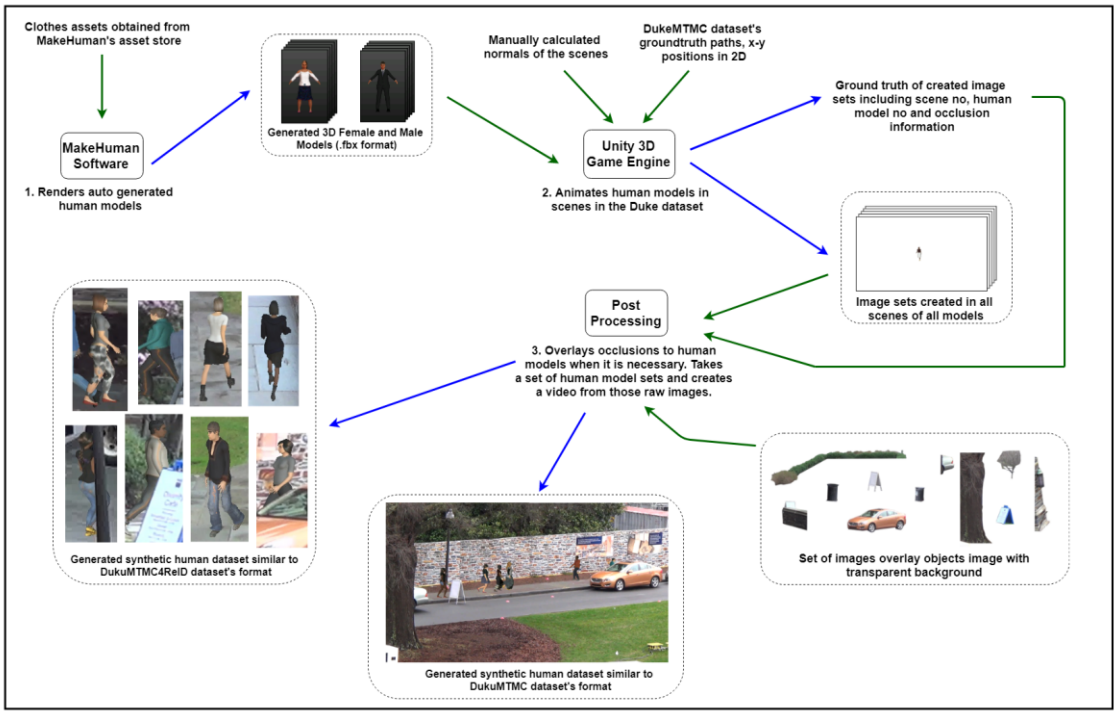


Figure 3.6: Overview: (1) get appropriate clothing assets from MakeHuman's Asset Store and create 560 synthetic male and female models using automation script written to generate human models for MakeHuman, then using MakeHuman convert these models to "fbx" format. (2) Unity imports these imports these generated 3D models and animates them according to the real dataset's ground truth data. While animating models, the scene slopes are considered to keep models oriented. (3) Image sets created by Unity are handled in Post Processing phase. In this phase, images of different models are brought together to create videos. Post Processing phase handles occlusions by overlaying transparent occlusion images to model images. Finally, according to selected real dataset, the format generated images are set and they are exported as either image sets or videos.

This phase starts with indexing the humanoids' images according to the ground truth data generated previously, so which humanoids will be shown in which frame is known. For each frame, the humanoid that needs to be active in this frame are extracted from their white background and they are overlaid on top of the active scene's empty background image. This overlay operation is done according to humanoids distance from the camera, the ones further away from the camera are overlaid earlier than the ones closer to the camera. In this way, not only occlusion objects occlude humanoids,

but also the humanoids occlude each other too. The distance to the camera is measured by the x position of the humanoids in 2dimage. Larger x values mean they are closer to the camera. Mean-while, every time a humanoid is overlaid, if its ground truth data contains any occlusion information, those transparent occlusion images are overlaid on top of the scene as well. When these operations are done for a frame, before it iterates to the next frame, the humanoids are cropped out from the whole image to create a DukeMTMC-VideoReID like dataset. Then these cropped images are grouped similar to the DukeMTMC-VideoReID dataset. At this point, instead of cropping out the images, the entire image can be saved as full hd video just like DukeMTMC [12] dataset as well.

Figure 3.7: 4 Example images from the training set of the DukeMTMC4ReID dataset. The first 2 images belong to the scene 1 and the other 2 belong to the scene 8. This person is included in only these 2 scenes.

Using this 3-phase system, we can create a highly automated synthetic data set that mimics the DukeMTMC-VideoReID dataset. In fact, at some point we are positively differentiating from the actual data set. For instance, in the real dataset the training data contains702 identities but only 2196 image sets, so each identity appear in3 cameras' viewpoint on average. Unlike the real dataset, in our synthetic dataset, each identity appears in all cameras' viewpoints. Thus, although we used 560 identities, the number of image sets in our synthetic dataset is twice the real dataset.

Figure 3.8: Occlusion objects in DukeMTMC-VideoReID dataset. These objects are used to oclude synthetic humanoids in the post processing phase.

4. RESULTS

The purpose of person re-identification is identifying the same person among people and from different cameras. Thus, a dataset which covers people from multiple cameras is a must. We generated this dataset synthetically by mimicking the DukeMTMC-VideoReIDdataset. In order to show the performance of our synthetic dataset, we used our dataset to train a CNN model (ETAP-Net) [34] which addresses the person re-identification. This paper presents different approaches for video-based person re-identification problem, supervised and semi-supervised methods. Besides ETAP-Net model, some tests are performed with the NVAN model [23] as well.

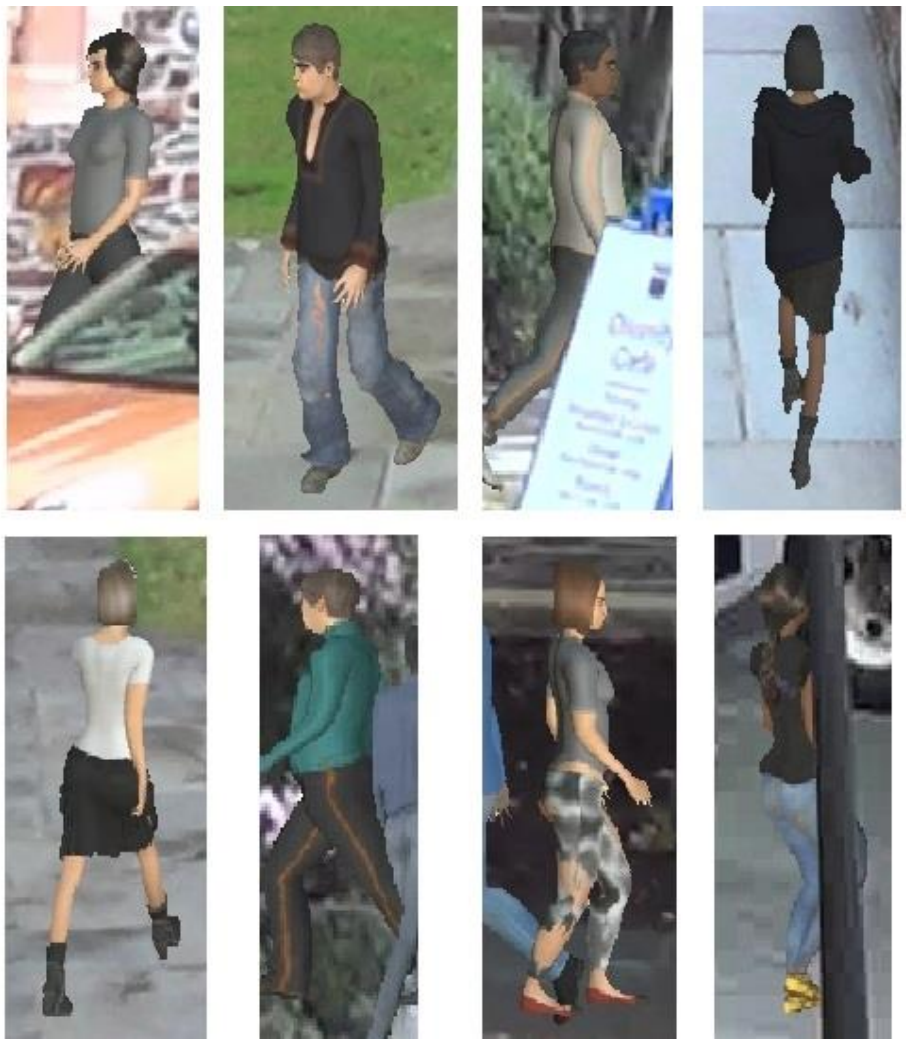


Figure 4.1: Synthetic images mimicking DukeMTMC-VideoReID dataset. Each image belongs to a different identity and a different scene. As seen here, image sizes may vary depending on the distance to the camera.

In order to evaluate our methods, we use the Cumulative Matching Characteristic (CMC) curve and the mean average precision (mAP). Since our target is boosting video based person re-identification, we are computing the average precision of each query from its precision-recall curve. In short, the videos provided for 702 test data are evaluated in the evaluation phase.

In Wu's paper [34], ETAP-Net is used as basic CNN model. They introduce a supervised video based person re-identification model. In the experiments, they take an ImageNet [18] pre-trained ResNet-50 model with last classification layer removed as the initialization of ETAP-Net. The experiments are done using PyTorch [26]. In the paper following parameters used: 70 epochs with a batch size of 16in each iteration, and overall learning rate is initialized to 0.2 and changed to 0.01 in the last 15 epochs.

In addition to supervised method, a semi-supervised approach that uses a progressive method for one-shot setting is introduced by Wu et al. The name of this proposed stepwise learning method is Exploit the Unknown Gradually (EUG). At first the training data is divided into 2 subgroups, labeled and unlabeled data. Exactly one tracklet obtained from each person in training data and placed into labeled data group. The rest of the training data considered as unlabeled data. Firstly, the labeled data is used as training data and tested on unlabeled data. Then training data is updated in every iteration with unlabeled data. At the end of every iteration, the trained data is used to predict unlabeled data. After this prediction, the unlabeled data is sorted by prediction reliability and the best-performing sub-set is selected (called pseudo labeled data) and moved from unlabeled data to labeled data. Therefore, in every iteration the training data is increased with selected unlabeled data. In the last iteration, when all unlabeled data is converted to pseudo labeled, the CNN model is trained with enlarged training data and tested on the test data. As noted in the paper, this progressive sampling method to increase the number of selected training candidates step by step outperforms the state-of-the-art methods on the DukeMTMC-VideoReID dataset. In this semi-supervised method, the tests are conducted with enlarging factors from 0.30 to 0.05 which indicates the speed of the sampled subset.

ETAP-Net Test Results with Real Dataset (702 people, 2196 tracklets)					
Enlarging Factor	Top-1	Top-5	Top-10	Top-20	mAP
0.30	38.5	51.7	56.3	63.4	32.8
0.20	46.6	61.0	66.8	72.9	40.6
Supervised	88.5	97.7	98.6	-	84.5

Table 4.1: Experiment results obtained with DukeMTMC-VideoReID dataset. Our evaluation metrics are Cumulative Matching Characteristic (CMC) curve and mean average precision (mAP). Due to limited memory in our GPU, we had to decrease the batch size from 16 to 4. In the bottom of the table, the one-shot stepwise learning method is ignored and entire training data are labeled and no iteration is made to increase the training data gradually. For supervised method there is no result for top-20.

As shown in Table 4.1, the results show that both CMC and mAP rates increases as the enlarging factor decreases. The decrease in the enlarging factor means an increase in the number of iterations in the training phase. An increase in the number of iterations means selecting less unlabeled data during each iteration. As a result, we concluded that gradual increase in training data appears to have a significant impact on success rates. However, the best accuracies for top-1, top-5, top-10, top-20 and mAP are obtained in fully supervised method. The results show that there is a significant gap between results obtained with the supervised and semi-supervised methods. However, the results also indicates that even if there are not enough labeled training data, the CNN model based on one-shot stepwise learning approach for person re-identification with sufficient amount of iteration can be utilized and achieve rather good results, but not as good as fully the supervised method.

In order to evaluate our synthetic dataset’s effect on person re-identification, we enlarged the DukeMTMC-VideoReID dataset with560 synthetic humans along with

4444 tracklets. Thereby, we increased the number of people from 702 to 1262 and number of tracklets from 2196 to 6640. We expected to see this large expansion of the dataset to have very effective results on success, and as we expected the success rates CMC and mAP rates are increased in line with our expectations. The results shown in Table 4.2 were obtained with ETAP-Net using the same parameters but different datasets than those shown in Table 4.1. As can be seen, the results obtained using our synthetic dataset is promising.

ETAP-Net Test Results with Mixed Datasets 702real + 560synthetic people, 2196real + 4444synthetic tracklets					
Enlarging Factor	Top-1	Top-5	Top-10	Top-20	mAP
0.30	59.7	59.7	59.7	59.7	59.7
0.20	72.8	72.8	72.8	72.8	72.8
Supervised	90.6	90.6	90.6	90.6	90.6

Table 4.2: Experiment results obtained by expanding the DukeMTMC-VideoReID dataset with synthetic data. The same parameters are used to get the results. Besides, the same evaluation metrics as in Table 4.1 are used to present gathered achievements. For supervised method there is no result for top-20.

To analyze the effect of increasing the dataset size with real and synthetic data on the success rate in more detail, we conducted new tests by gradually increasing the number of identities in the training dataset. As can be seen in following tables, increasing the number of identities of both supervised and semi-supervised methods have a positive effect on the success rate.

4.1. ETAP-Net Supervised Method

In this section, the real, synthetic and mixed datasets' contributions to the accuracy of the person re-identification will be examined. The results in this section were produced using Wu et al.'s supervised ETAP-Net model.

ETAP-Net Supervised Method Test Results with Real Dataset (702 people, 2196 tracklets)				
Number of Real Identity	Top-1	Top-5	Top-10	mAP
102	62.8	62.8	62.8	62.8
202	70.4	70.4	70.4	70.4
302	76.5	76.5	76.5	76.5
402	79.6	79.6	79.6	79.6
502	80.6	80.6	80.6	80.6
602	84.2	84.2	84.2	84.2
702	88.5	88.5	88.5	88.5

Table 4.3: Experiment results obtained with DukeMTMC-VideoReID dataset using supervised method. Our evaluation metrics are Cumulative Matching Characteristic (CMC) curve and mean average precision (mAP). For this experiment, batch size is set to 8.

The results in Table 4.3 and 4.4 indicate that the performance of the Neural Network algorithms depends on the size of the dataset. This correlation can be seen from both real and synthetic datasets. As can be seen in Table 4.3, the top-1 accuracy can be increased up to 41%by increasing the number of training identity. Similar results can be seen in Table 4.3 with synthetic data as well. The top-1 accuracy increased up to 22% by increasing the number of synthetic identity. In both tests, same parameters used and

the only difference is the type of the training dataset; however, the outcome is same for both, the significance of the dataset size in Neural Network.

ETAP-Net Supervised Method Test Results with Synthetic Dataset (500 people, 4000 tracklets)				
Number of Synthetic Identity	Top-1	Top-5	Top-10	mAP
100	32.9	32.9	32.9	32.9
200	41.5	41.5	41.5	41.5
300	38.7	38.7	38.7	38.7
400	37.9	37.9	37.9	37.9
500	40.2	40.2	40.2	40.2

Table 4.4: Experiment results obtained with synthetic dataset using super-vised method. Our evaluation metrics are Cumulative Matching Characteristic (CMC) curve and mean average precision (mAP). For this experiment, batch size is set to 8.

Table 4.3 and 4.4 show the impact of the dataset size on accuracy, however it is clear that the performance of synthetic data only approach is lower than the real data used approach. Table 4.5 shows the case where there is a small amount of real identity and large amount of synthetic identity. This experiment shows that even if the real dataset's performance cannot be over passed by synthetic data, the synthetic data still can be used to boost performance of the small dataset. The results of the expanded synthetic data with a small amount of real data are quite elevated.

ETAP-Net Supervised Method Test Results with Mixed Datasets 100 real and up to 500 Synthetic Identity				
Number of Synthetic Identity	Top-1	Top-5	Top-10	mAP
100	67.9	67.9	67.9	67.9
200	68.5	68.5	68.5	68.5
300	70.4	70.4	70.4	70.4
400	72.9	72.9	72.9	72.9
500	70.7	70.7	70.7	70.7

Table 4:5: Experiment results obtained with mixed datasets using supervised method. In each test 100 real identities are used and along with a number of synthetic data. The number of humanoids increased gradually up to 500.Our evaluation metrics are Cumulative Matching Characteristic (CMC) curve and mean average precision (mAP). For this experiment, batch size is set to 8.

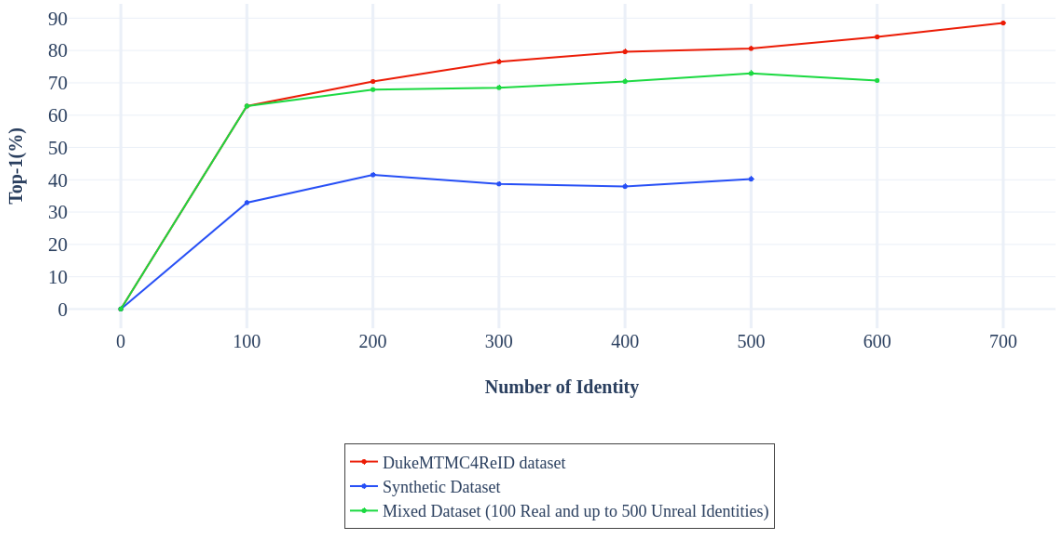


Figure 4.2: Comparisons of different datasets using Wu's supervised method[34] with batch size of 8. DukeMTMC-VideoReID (real), synthetic and partially synthetic datasets are tested.

Figure 4.2 emphasizes the significance of the size of the dataset regardless of the type of data it contains. As can be seen in the figure, as the number of identities increases, the top-1 results increase with respect to the dataset size as well. Another argument to be drawn from this plot, even if the success of the synthetic dataset alone is not sufficient, high results can be achieved by adding a small amount of real data.

4.2. ETAP-Net Semi-supervised Method (EUG)

The results obtained by Wu's convolutional neural network approach, stepwise progressive learning model, for person re-identification [34], show that boosting the training dataset is an important factofor better result. We boosted the training dataset of DukeMTMC-VideoReID with synthetically generated photo realistic human agents and results show that we achieved considerable success.

ETAP-Net Semi-supervised Method Test Results with Real Dataset (702 people, 2196 tracklets)					
Number of Real Identity	Top-1	Top-5	Top-10	Top-20	mAP
102	35.2	35.2	35.2	35.2	35.2
202	14.5	14.5	14.5	14.5	14.5
302	40.5	40.5	40.5	40.5	40.5
402	50.4	50.4	50.4	50.4	50.4
502	47.9	47.9	47.9	47.9	47.9
602	43.3	43.3	43.3	43.3	43.3
702	46.6	46.6	46.6	46.6	46.6

Table 4.6: Experiment results obtained with DukeMTMC-VideoReID datasetusing semi-supervised method. Our evaluation metrics are CumulativeMatching Characteristic (CMC) curve and mean average precision (mAP).For this experiment, batch size is 4 and enlarging factor is set to 0.20.

The results shown in Table 4.1 and 4.2 are obtained using the same input parameters and in the same environment. As seen, the top-1accuracy increased from 38.5% to 59.7% with 0.30 enlarging factor and from 46.6% to 72.8% with 0.20 enlarging factor. In addition to that, the mean average precision (mAP) rates are increased from32.8% to 52.7% and 40.6% to 65.4% with 0.30 and 0.20 enlarging factors respectively. These rate of changes prove how successful our synthetic dataset is and can be utilized to boost a person re-identification dataset.

We also repeated the tests we performed in Table 4.3 and 4.4 for the ETAP-Net Semi-supervised method (EUG). To analyze the effect of increasing the dataset with real and synthetic data on success rate, we conducted new tests by gradually increasing the number of training identity in the dataset. As can be seen in Table 4.6 and 4.7, increasing the number of both real and synthetic identities of semi-supervised method have positive effects on the success rate as well in the supervised approach. Although we have improved the success rates for both datasets in our experiments, as can be seen in the tables, the results obtained by the increase in the real dataset are higher than the synthetic dataset.

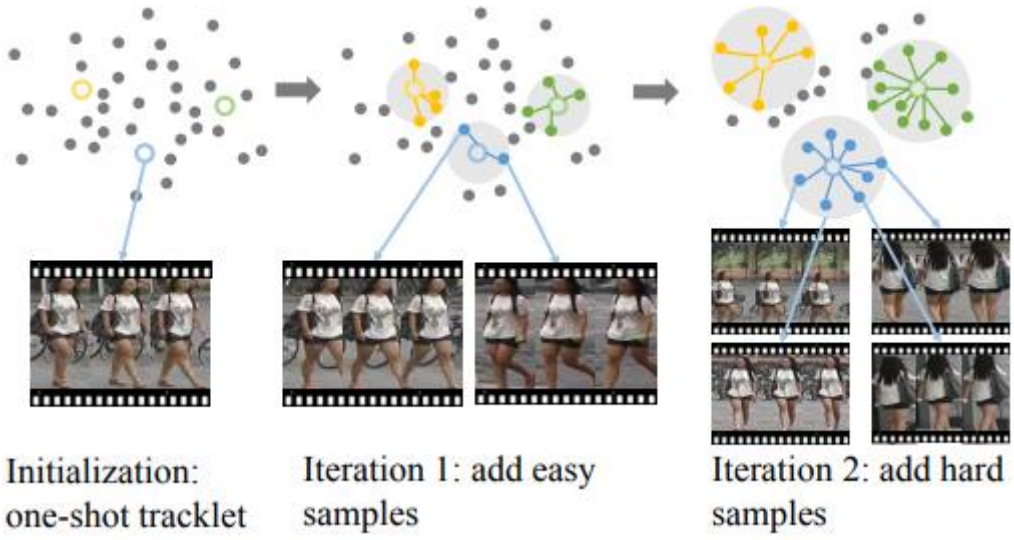


Figure. 4.3: An illustration of the ETAP-Net semi-supervised method's unlabeled data sampling procedure.

ETAP-Net Semi-supervised Method Test Results with Synthetic Dataset (500 people, 4000 tracklets)					
Number of Synthetic Identity	Top-1	Top-5	Top-10	Top-20	mAP
100	15.4	15.4	15.4	15.4	15.4
200	13.4	13.4	13.4	13.4	13.4
300	17.9	17.9	17.9	17.9	17.9
400	18.8	18.8	18.8	18.8	18.8
500	15.7	15.7	15.7	15.7	15.7

Table 4.7: Experiment results obtained with synthetic dataset using semi-supervised method. Our evaluation metrics are Cumulative Matching Characteristic (CMC) curve and mean average precision (mAP). For this experiment, batch size is 4 and enlarging factor is set to 0.20.

In order to analyze the effect of increasing the dataset size with synthetic data on success rate, we conducted new tests by gradually increasing the number of humanoids in the synthetic dataset. As can be seen in both Table 4.4 and 4.7, increasing the number of humanoids of both supervised and semi-supervised methods has significant effect on the success rate. Although the results of these tests are lower than the ones conducted with the real dataset, the positive contribution of the expansion of the dataset to the success rate appears in both cases. Based on this argument, for semi-supervised method the tests are repeated with a partially mixed dataset as done in Table 4.5. In this case, it is simulated that there is a small number of real identities in the dataset and that they are supported by a large number of synthetic data. This data used to train ETAP-Net semi-supervised method. The test results are listed in Table 4.8, and the success rate is significantly increased compared to the case where entire dataset is composed of synthetic data. These results are very similar to the results obtained supervised method.

ETAP-Net Semi-supervised Method Test Results with Mixed Dataset 100 Real and up to 500 Synthetic Identity					
Number of Synthetic Identity	Top-1	Top-5	Top-10	Top-20	mAP
100	53.4	70.5	74.6	79.5	46.0
200	55.0	71.5	77.4	82.6	47.5
300	54.4	71.5	77.6	83.0	46.7
400	51.7	69.8	75.8	81.3	45.5
500	45.4	62.8	70.7	78.1	40.2

Table 4.8: Experiment results obtained with mixed datasets using semi-supervised method. In each test 100 real identity is used and along with number of synthetic data. Number of humanoid increased gradually up to500. Our evaluation metrics are Cumulative Matching Characteristic (CMC) curve and mean average precision (mAP). For this experiment, batch size is4 and enlarging factor is set to 0.20.

4.3. NVAN Supervised Method

In order to show that the success achieved with our synthetic dataset is not solely dependent on Wu's semi-supervised and super-vised methods [34], several tests are performed by another training method using the similar approaches used in previous experiments. The new tests performed using Liu et al.'s NVAN model [23]. NVAN is a video based person re-identification model build on Pytorch. In this work, Liu et al. target at refining the intermediate features as well as high-level features with non-local attention operations and **t**hey propose this Non-local Video Attention Network (NVAN) to incorporate video characteristics into the representation at multiple feature levels.

Using NVAN model 3 tests conducted with following dataset con-figurations; DukeMTMC-VideoReID dataset (entirely real identities), synthetic dataset (entirely synthetic identities), and mixed dataset (combination of entire real and synthetic identities). Due to GPU memory issues, same parameters offered by Liu et al. could not be used, but similar results obtained with some lower parameters. The results of the tests are shown in Table 4.9.

As can be seen from the Table 4.9, although the synthetic dataset alone does not yield enough accuracy, merging synthetic and real identities achieves very good compared to DukeMTMC-VideoReID dataset. Using mixed dataset instead of DukeMTMC-VideoReID dataset the accuracy of top-1 ranking is increased from 94.6% to95.6% and mAP is increased from 91.5% to 93.1%. That is, our mixed dataset approach outperforms the real dataset only approach with NVAN model as it was with ETAP-Net model.

NVAN Supervised Method		
Training Dataset	Top-1	mAP
Synthetic	22.7	17.7
DukeMTMC-VideoReID	94.6	91.5
Mixed	95.6	93.1

Table 4.9: Comparisons of different datasets using NVAN supervised method [23]. DukeMTMC-VideoReID (real), synthetic and mixed datasets are tested.

4.4. Comparison with the State-of-art Approaches

Comparison of the state-of-art Reports On the DukeMTMC-VideoReID Dataset		
Method	Top-1	mAP
Supervised [32]	87.0	83.5
Active Learn [32]	85.2	80.1
BUC [22]	74.8	66.7
OIM [35]	51.1	43.8
GAN [5]	82.2	78.8
DCML [33]	79.2	75.3
Baseline STA [11]	79.1	76.8
Baseline + TL + STA + Fusion + Reg [11]	96.2	94.9
STCnet	95.0	93.5
GLTR	96.3	93.8
Stepwise [34]	56.3	46.8
ETAP-Net (one shot) [34]	39.7	33.3
ETAP-Net* + semi-supervised [34]	69.0	59.5
ETAP-Net* + supervised [34]	83.6	78.3
ETAP-Net** + semi-supervised [34]	46.6	40.6
ETAP-Net** + supervised [34]	88.5	84.5

Ours + ETAP-Net + semi-supervised	**72.8**	**65.4**
Ours + ETAP-Net + supervised	**90.6**	**87.7**
NVAN* [23]	96.3	94.9
NVAN** [23]	94.6	91.5
Ours + NVAN* [23]	**95.6**	**93.1**

Table 4.10: Comparison of our methods (fully and semi supervised) with state-of-the-art methods on the DukeMTMC-VideoReID dataset. In the table, multiple ETAP-Net supervised and semi-supervised results are listed. Results marked as ETAP-Net* express the reports shared in Wu et al. [34].The results marked as ETAP-Net** on the other hand, expresses the results we obtained with the same method and dataset in our test environment. The reason for the differences between these results is the parameters we use. Due to our hardware differences, we had to change some parameters differently such as max_frames, batch_size etc. to overcome the memory problem. The same applies to the NVAN model NVAN* expresses the results hared in Liu et al.[23] and NVAN** expresses the results we obtained with the same model and dataset in our test environment. Due to lack of memory issue, we decreased the batch size to half and this created the difference between our results with the one shared in the paper.

In Table 4.10, the state-of-art methods' reports on DukeMTMC-VideoReID dataset are listed. The table also includes the base methods we have used in this paper, EUG supervised and semi-supervised methods and NVAN. The ours keyword in the table denotes that the DukeMTMC-VideoReID dataset is mixed with our synthetic dataset and in this way the dataset size is increased from 702 to1262 identities. Then this enlarged dataset used to train EUG and NVAN methods and as shown in the table our synthetically boosting approach outperformed the base methods' results.

Our approach not only outperformed the base methods' result but also other stat-of-art methods' results as well. Our ETAP-Net semi-supervised and supervised approaches achieved 72.8%, 90.6% top-1 rankings respectively. In addition to that, our NVAN mixed dataset approach achieved 95.6% top-1 ranking, and have made a significant difference to their state-of-art competitors.

5. CONCLUSION

In order to get satisfying results from convolutional neural net-work models for person re-identification, it is essential to provide a proper and large dataset to it. As long as the dataset provides more data and diversity, its positive effect on the success rate of the CNN model becomes more pronounced. To meet this need, we propose a synthetic dataset to boost the real dataset. We utilized DukeMTMC-VideoReID Dataset and boosted it with our synthetic dataset. Training and testing operations of the real and mixed datasets performed on the same CNN based models ETAP-Net [34] and NVAN [23]. ETAP-Net provides stepwise progressive learning method in addition to supervised method. We collected results from theses CNN models for both semi-supervised and fully-supervised methods. As a result, for ETAP-Net semi-supervised method the top-1 accuracy of our mixed dataset approach outperformed the real dataset used method by26.2%, in terms of top-1 accuracy we achieved 72.8%. For ETAP-Net fully-supervised method we get similar CMC results in experiments, 88.5% and 90.6% top-1 accuracy with real and mixed datasets respectively. In addition to this, we achieved some increase in mean average precision too. The mean average precision increased from84.5% to 87.7% using synthetically boosted data. Therefore, we out-performed our base EUG model's both fully and semi supervised results. In order to prove that our approach is not entirely dependent on ETAP-Net model, we conducted some tests with NVAN model as well. The results obtained from this method bear resemblance to those from the ETAP-Net. As in our tests with the ETAP-Net model, we achieved a significant increase with mixed dataset. The top-1 accuracy of mixed dataset approach outperformed the real dataset approach by 1.0% and mAP outperformed by 1.6%. Thus, our mixed dataset approach has been successful in the NVAN as in the ETAP-Net methods. Our accomplishment goes beyond these two methods and stands out among the results of the state-of-art methods listed in Table 4.10.

To conclude, we propose a framework to generate photo-realistic synthetic data based on DukeMTMC-VideoReID dataset. Our almost fully automated dataset generator framework can be utilized to create data to train a CNN model for video based person re-identification. After manually setting some parameters, our framework can create the dataset without any human contribution. Thus, even though the desired background or

scene of the dataset changes, with limited manual modifications, a new synthetic dataset can be created easily.

5.1. Future Work

In the future, some manual parameters of the framework (such as; walking paths, ground normals, animation speeds etc.) can be obtained automatically using computer vision techniques. In addition to this, automatically generated agents can be optimized and diversified by using more clothing.

6. BIBLIOGRAPHY

[1] Carnegie Mellon University - CMU Graphics Lab - motion capture library, (n.d.). http://mocap.cs.cmu.edu (Accessed June 10, 2019).

[2] MakeHuman Software, (n.d.). http://www.makehumancommunity.org (Accessed June 10, 2019).

[3] Unity Software, (n.d.). https://unity.com (Accessed June 10, 2019).

[4] B. Benfold, I. Reid, Stable multi-target tracking in real-time surveillance video, in: CVPR 2011, 2011: pp. 3457–3464.

[5] A. Borgia, Y. Hua, E. Kodirov, N. Robertson, GAN-based Pose-aware Regulation for Video-based Person Re-identification, in: 2019 IEEE Winter Conf. Appl. Comput. Vis., 2019: pp. 1175–1184.

[6] L. Briceno, G. Paul, MakeHuman: A Review of the Modelling Framework: Volume V: Human Simulation and Virtual Environments, Work With Computing Systems (WWCS), Process Control, in: 2019: pp. 224–232. doi:10.1007/978-3-319-96077-7_23.

[7] W. Chen, H. Wang, Y. Li, H. Su, Z. Wang, C. Tu, D. Lischinski, D. Cohen-Or, B. Chen, Synthesizing training images for boosting human 3d pose estimation, in: 2016 Fourth Int. Conf. 3D Vis., 2016: pp. 479–488.

[8] E. Cheung, A. Wong, A. Bera, D. Manocha, Mixedpeds: pedestrian detection in unannotated videos using synthetically generated human-agents for training, in: Thirty-Second AAAI Conf. Artif. Intell., 2018.

[9] E. Cheung, T.K. Wong, A. Bera, X. Wang, D. Manocha, Lcrowdv: Generating labeled videos for simulation-based crowd behavior learning, in: Eur. Conf. Comput. Vis., 2016: pp. 709–727.

[10] M.A. Fischler, R.C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Commun. ACM. 24 (1981) 381–395.

[11] Y. Fu, X. Wang, Y. Wei, T. Huang, STA: Spatial-Temporal Attention for Large-Scale Video-based Person Re-Identification, in: Proc. Assoc. Adv. Artif. Intell., 2019.

[12] M. Gou, S. Karanam, W. Liu, O. Camps, R.J. Radke, DukeMTMC4ReID: A large-scale multi-camera person re-identification dataset, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Work., 2017: pp. 10–19.

[13] M. Gou, Z. Wu, A. Rates-Borras, O. Camps, R.J. Radke, others, A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2018) 523–536.

[14] H. Hattori, V. Naresh Boddeti, K.M. Kitani, T. Kanade, Learning scene-specific pedestrian detectors without real data, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015: pp. 3819–3827.

[15] K. Hattori, H. Hattori, Y. Ono, K. Nishino, M. Itoh, V.N. Boddeti, T. Kanade, Carnegie Mellon University Surveillance Research Dataset (CMUSRD), Tech. Report, Carnegie Mellon Univ. (2014).

[16] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, X. Chen, VRSTC: Occlusion-Free Video Person Re-Identification, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019: pp. 7183–7192.

[17] S. Huang, D. Ramanan, Expecting the unexpected: Training detectors for unusual pedestrians with adversarial imposters, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017: pp. 2243–2252.

[18] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Adv. Neural Inf. Process. Syst., 2012: pp. 1097–1105.

[19] A. Lerer, S. Gross, R. Fergus, Learning physical intuition of block towers by example, ArXiv Prepr. ArXiv1603.01312. (2016).

[20] J. Li, J. Wang, Q. Tian, W. Gao, S. Zhang, Global-Local Temporal Representations For Video Person Re-Identification, ArXiv Prepr. ArXiv1908.10049. (2019).

[21] W. Li, R. Zhao, T. Xiao, X. Wang, Deepreid: Deep filter pairing neural network for person re-identification, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014: pp. 152–159.

[22] Y. Lin, X. Dong, L. Zheng, Y. Yan, Y. Yang, A bottom-up clustering approach to unsupervised person re-identification, in: Proc. AAAI Conf. Artif. Intell., 2019: pp. 1–8.

[23] C.-T. Liu, C.-W. Wu, Y.-C.F. Wang, S.-Y. Chien, Spatially and Temporally Efficient Non-local Attention Network for Video-based Person Re-Identification, ArXiv Prepr. ArXiv1908.01683. (2019).

[24] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, M.J. Black, SMPL: A skinned multi-person linear model, ACM Trans. Graph. 34 (2015) 248.

[25] J. Marin, D. Vázquez, D. Gerónimo, A.M. López, Learning appearance in virtual scenarios for pedestrian detection, in: 2010 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2010: pp. 137–144.

[26] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in pytorch, (2017).

[27] H. Rahmani, A. Mian, M. Shah, Learning a deep model for human action recognition from novel viewpoints, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2017) 667–681.

[28] E. Ristani, F. Solera, R. Zou, R. Cucchiara, C. Tomasi, Performance measures and a data set for multi-target, multi-camera tracking, in: Eur. Conf. Comput. Vis., 2016: pp. 17–35.

[29] K.M. Robinette, S. Blackwell, H. Daanen, M. Boehmer, S. Fleming, Civilian American and European Surface Anthropometry Resource (CAESAR), Final Report. Volume 1. Summary, 2002.

[30] D. Thirde, L. Li, F. Ferryman, Overview of the PETS2006 challenge, in: Proc. 9th IEEE Int. Work. Perform. Eval. Track. Surveill. (PETS 2006), 2006: pp. 47–50.

[31] G. Varol, J. Romero, X. Martin, N. Mahmood, M.J. Black, I. Laptev, C. Schmid, Learning from synthetic humans, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017: pp. 109–117.

[32] M. Wang, B. Lai, Z. Jin, X. Gong, J. Huang, X. Hua, Deep Active Learning for Video-based Person Re-identification, ArXiv Prepr. ArXiv1812.05785. (2018).

[33] N. Wojke, A. Bewley, Deep cosine metric learning for person re-identification, in: 2018 IEEE Winter Conf. Appl. Comput. Vis., 2018: pp. 748–756.

[34] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, Y. Yang, Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018: pp. 5177–5186.

[35] T. Xiao, S. Li, B. Wang, L. Lin, X. Wang, Joint detection and identification feature learning for person search, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017: pp. 3415–3424.

[36] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in: Proc. IEEE Int. Conf. Comput. Vis., 2015: pp. 1116–1124.

HACETTEPE UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING
THESIS/DISSERTATION ORIGINALITY REPORT

HACETTEPE UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING
TO THE DEPARTMENT OF Computer Engineering

Date: 18/09/2019

Thesis Title / Topic: Boosting Video-Based Person Re-Identification with Synthetic Human Agents

According to the originality report obtained by my thesis advisor by using the *Turnitin* plagiarism detection software and by applying the filtering options stated below on 18/09/2019 for the total of 42 pages including the a) Title Page, b) Introduction, c) Main Chapters, d) Conclusion sections of my thesis entitled as above, the similarity index of my thesis is 10%.

Filtering options applied:

1. Bibliography/Works Cited excluded
2. Quotes excluded
3. Match size up to 5 words excluded

I declare that I have carefully read Hacettepe University Graduate School of Sciene and Engineering Guidelines for Obtaining and Using Thesis Originality Reports; that according to the maximum similarity index values specified in the Guidelines, my thesis does not include any form of plagiarism; that in any future detection of possible infringement of the regulations I accept all legal responsibility; and that all the information I have provided is correct to the best of my knowledge.

I respectfully submit this for approval.

Date and Signature

Name Surname: Fikret Kaya — 18/09/2019

Student No: N16125433

Department: Computer Engineering

Program:

Status: ☒ Masters ☐ Ph.D. ☐ Integrated Ph.D.

ADVISOR APPROVAL

APPROVED.

Dr.Öğr.Üyesi Ufuk Çelikcan

(Title, Name Surname, Signature)

CURRICULUM VITAE

Name Surname	: Fikret KAYA
Place of Birth	: Kulu
Date of Birth	: 01.05.1993
Marital Status	: Single
Address	: Ankara/Turkey
Phone	: +90 505 5685876
Email	: fikret.ky93@gmail.com
Foreign Languages	: English

EDUCATION

High School	: Gazi Anatolian High School, 2007-2011
Bsc.	: Bilkent University, Computer Science, 2011-2016

Work Experience

Turkish Aerospace	: April 2017 – March 2019
Havelsan	: March 2019 -