The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-2/W3-2023
ISPRS Intl. Workshop "Photogrammetric and computer vision techniques for environmental and infraStructure monitoring, Biometrics and Biomedicine"
PSBB23, 24–26 April 2023, Moscow, Russia

MULTIMODAL PERSON RE-IDENTIFICATION IN AERIAL IMAGERY BASED ON CONDITIONAL ADVERSARIAL NETWORKS

V.V. Kniaz[1,2], V.A. Knyaz[1,2], P.V. Moshkantsev[1]

[1] State Res. Institute of Aviation Systems (GosNIIAS), 125319, 7, Victorenko str., Moscow, Russia
(vl.kniaz,knyaz, petr_mosh,)@gosniias.ru
[2] Moscow Institute of Physics and Technology (MIPT), Russia

Commission II, WG II/8

KEY WORDS: person re-identification, generative adversarial networks, thermal images, airborne images.

ABSTRACT:

Person Re-Identification (Re-ID) is the task of matching the same person in multiple images captured by different cameras. Recently deep learning-based Re-ID algorithms demonstrated an exciting progress for terrestrial-based cameras, still, person Re-ID in aerial images poses multiple challenges including occlusion of human feature parts, image distortion, and dynamic camera location. In this paper, we propose a new Person Aerial Re-ID framework Robust to Occlusion and Thermal imagery (`ParrotGAN`). Our model is focused on cross-modality person Re-ID in aerial images. Furthermore, we collected a new large-scale synthetic multimodal *AerialReID* dataset with 30k images and 137 person ID. Our `ParrotGAN` model leverages two strategies to achieve robust performance in the task of person Re-ID in thermal and visible range. Firstly, we use a latent space of `StyleGAN2` model to estimate the distance between two images of a person. Specifically, we project each real image into the latent space with a correspondent latent vector z. We use the distance between latent vectors to provide a Re-ID similarity metric. Secondly, we use a generative-adversarial network to translate a color image to a synthetic thermal image. We use synthetic image for a cross-modality Re-ID. We evaluate our `ParrotGAN` model and basleines on our *AerialReID* and *PRAI-1581* datasets. The results of the evaluation are encouraging and demonstrate that our `ParrotGAN` model competes with baselines in visible range aerial person Re-ID and outperforms them in the cross-modality setting. We made our code and dataset publicly available.

1. INTRODUCTION

Person Re-Identification (Re-ID) is the task of matching the same person in multiple images captured from different viewpoints. The input for a Re-ID method includes two sets of images. The probe set includes one or more images of the person that must be identified in the new environment. The gallery set includes images that may contain the person from the probe image. The Re-ID task's complexity arises from the differences in the person appearance in the probe and gallery sets. Person appearance may be different in different cameras due to changes in illumination or viewpoint locations.

The task of person Re-ID (Nguyen et al., 2017, Nguyen and Park, 2016, Ye et al., 2018b, Ye et al., 2018a) received a lot of scholar attention recently. Such methods can be broadly divided into three groups: direct methods, metric learning, and deep learning. Direct methods aim to detect discriminative features in person appearance that can be matched robustly among different cameras. Metric learning methods provide a function that returns a distance for a given pair of samples in the probe and gallery sets. The distance is required to be small if the pair is correct and large otherwise. Deep learning methods employ deep neural networks to learn end-to-end models for matching objects in the probe and gallery sets.

While many solutions have been proposed for the Re-ID task for images captured in the visible range, cross-modality ReID remains challenging. Recently a new generation of neural networks has been developed focusing on generative learning. Such networks are commonly called Generative Adversarial Networks (GANs) (Goodfellow et al., 2014). GANs are capable of learning complex image-to-image translations such as a season change or an object transfiguration. Modern research demonstrates that GANs can learn to translate probe images to different viewpoints or different illumination conditions. To the best of our knowledge, there are no results to date in the literature regarding cross-modality object ReID from airborne images.

This paper focuses on developing a deep learning `ParrotGAN` framework for cross-modality person Re-ID in thermal images. We use assumptions made by Kniaz et. al (Kniaz et al., 2019) as a starting point for our research. Our aim is to develop a model that can find matching person IDs in thermal gallery set given an input color probe image. To achieve this, we leverage two generative modelling techniques.

Firstly, we use a `ThermalGAN` (Kniaz et al., 2019) model to perform multimodal color-to-thermal image translation. Secondly, we develop a new Re-ID matching approach based on the projection to `StyleGAN2` (Karras et al., 2020b) latent space. Specifically, we develop a modified version of a `StyleGAN2` with two main contributions: (1) a modified discriminator that operates with pairs of person IDs and predicts their similarity, (2) a new loss function that aims to provide disengagement of the latent space for the task of person Re-ID. Given two person ID images A^1, A^2, we project them to two latent codes w_1, w_2. We use the distance in latent space to define the Re-ID similarity distance.

We developed a new *AerialReID* synthetic dataset to train and validate our framework. Our dataset includes 30k pairs of color

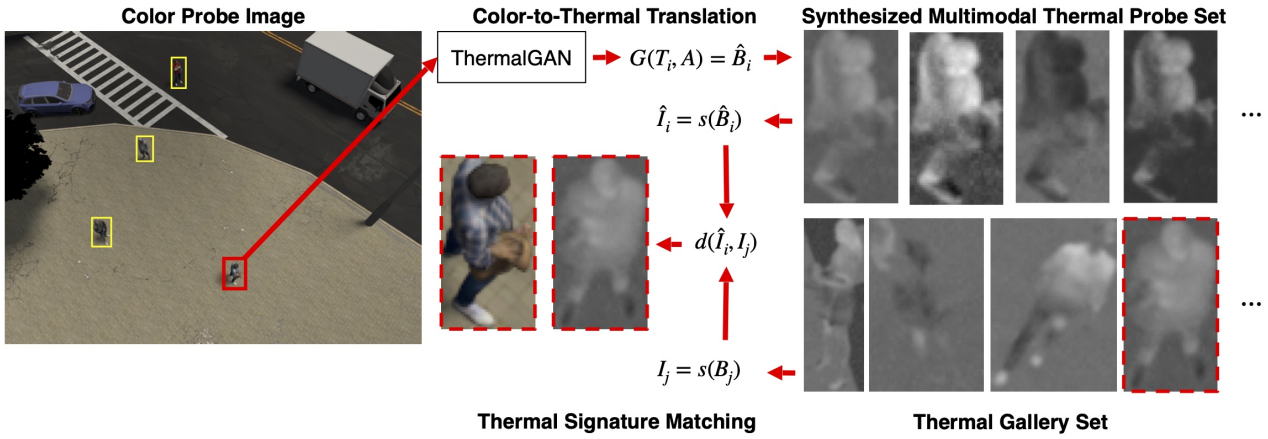

Figure 1. Overview of our proposed ParrotGAN framework.

and thermal images of 137 person IDs. We trained our framework using the training split of our dataset *AerialReID*dataset and validated it using the test split and samples from the *PRAI-1581* dataset (Zhang et al., 2021). The results of the evaluation are encouraging and demonstrate that our ParrotGAN model competes with
baselines in visible range aerial person Re-ID and outperforms them in the cross-modality setting.

1.1 Contributions

We present four key technical contributions:

- A new ParrotGAN model for cross-modality person Re-ID in aerial images using latent space projection.
- A new Re-ID distance metric based on image back-projection into a latent space of StyleGAN2 model.
- A new aerial person Re-ID dataset *AerialReID* with 30k pairs of thermal and visible range images and 137 person IDs.[1]
- An evaluation of our ParrotGAN model and five modern baselines on our *AerialReID* and *PRAI-1581* datasets.

2. RELATED WORK

2.1 Person Re-Identification

The problem of person re-identification is important for various computer vision applications, such multi-modal image segmentation and object detection, autonomous driving, security etc. So currently it attracts attention of many researches (Farenzena et al., 2010a, Gong et al., 2014, Wu et al., 2017, Wang et al., 2019, Kniaz and Knyaz, 2019, Bhuiyan et al., 2018, Prosser et al., 2008a, Bhuiyan et al., 2015).

2.1.1 Direct Methods Bird et al. (Bird et al., 2005) proposed a method based on subdividing the person in horizontal stripes, and keeping the median color of each stripe accumulated over different frames. In (GHEISSARI, 2006), a spatio–temporal local-feature-grouping and matching is proposed by building decomposable triangular graph in order to capture the spatial distribution of the local descriptor over time. The method proposed in (Wang et al., 2007) segments a pedestrian image from similar perspectives, and stores the spatial relationship of the colors into a co-occurrence matrix. To describe human bodies speeded up robust features (SURF) interest points are collected over short video sequences in (Hamdoun et al., 2008). A sophisticated appearance model, the symmetry-driven accumulation of local features (SDALF) has found application in work (Farenzena et al., 2010b). It models human appearance by the symmetry and asymmetry driven features, based on the idea that features closer to the symmetry axes are more robust against scene clutter than are the features far away from them. In the same pipeline, (Cheng et al., 2011) proposed a pictorial structure based (PS-based) ReID method by matching signature features with help of a number of well-localized body parts and manual weighing of these parts based on their conspicuity. Developing this idea (Bhuiyan et al., 2014) devised a method that automatically appoints weights to the body parts on the basis of their distinctive power. Covariance features, originally employed for pedestrian detection (Tuzel et al., 2008), are extracted from coarsely located body parts and adapted and modernized for ReID purposes in (Bak et al., 2010).

2.1.2 Metric Learning Metric Learning involves using the techniques that use the training data to learn spaces, in order to guarantee high re-identification rates. The main idea is that the learned information from the training data could be generalized to the unseen probe. The approach presented in (Gray and Tao, 2008) uses gain to select a combination of spatial and color information for invariance point of view. On the same principle, in (Schwartz and Davis, 2009), a multidimensional signature consisting of several functions is projected into a low-dimensional hidden discriminant space by Partial Least Squares (PLS) reduction. Re-identification is seen as a binary classification problem (one vs. all) by (Bak et al., 2010) using Haar-like functions and a part-based MPEG7 dominant color descriptor, while in (Prosser et al., 2008b) the relative ranking problem is presented in a higher dimensional feature space where true and false coincidence becomes more separable. In (Li et al., 2013), the adaptive boundary approach is proposed that simultaneously studies the metric and the locally adaptive thresholding rule. In (Zheng et al., 2012), a probabilistic relative distance comparison (PRDC) model is proposed that attempts to maximize the probability of a true match with a smaller distance than a false match. In (Koestinger et al., 2012), the statistical inference perspective is used to study a metric that satisfies equivalence constraints. In (Pedagadi et al., 2013), a local

[1] http://www.zefirus.org/ParrotGAN/

Fisherman discriminant analysis (LFDA) is proposed, which is a closed-form solution to the Mahalanobis matrix required to use principal component analysis (PCA) for dimensionality reduction. Inspite of this, PCA can eliminate discriminant functions, negating the benefits of the LFDA. By introducing the term regularization into methods (Mignon and Jurie, 2012, Li et al., 2013, Zheng et al., 2011) and using a series of kernel-based methods to study nonlinear feature transformation functions in (Pedagadi et al., 2013) to preserve distinctive features the study (Xiong et al., 2014) reported better performance compared to the corresponding methods.

Recently, deep-learning techniques have shown significant performance improvements in various computer vision applications such as image classification, object detection, and face recognition. Also, in recent years, the number of methods that apply deep learning methods (Chen et al., 2016, Varior et al., 2016a, Ahmed et al., 2015, Xiao et al., 2016, Varior et al., 2016b, Yan et al., 2016) for human ReID has been growing.

In (Chen et al., 2016, Varior et al., 2016a, Ahmed et al., 2015), a deep Siamese architecture for human ReID is proposed, which typically uses two or three Siamese convolutional neural networks (S-CNNs) for deep function learning. In (Ahmed et al., 2015), both feature representation and metric training were treated as a collaborative learning problem using an end-to-end deep convolutional architecture. It basically introduces a new deep architecture cross-entry neighborhood layer that efficiently extracts cross-functional relationships. A domain-guided dropout (DGD) (Xiao et al., 2016) studies robust functions by selecting neurons specific to certain domains.

2.1.3 Deep Learning An automatic search for a CNN architecture, proposed in (Quan et al., 2019), specifically suited to the reID task based on a search in a specially designed reID search space called Auto-ReID. Their Auto-ReID enables an automated approach to finding an efficient and effective CNN architecture for reID. In article (Wang and Zhang, 2020), ReID for an out-of-control person is formulated as a multi-label classification problem to gradually find the true labels. Method starts by assigning a single class of label to each human image and then moves on to a multi-label classification using the updated ReID model to predict the label. To solve problem of poor separation of two distance distributions for a positive sample of a pair (Pos-distr) and a pair of negative samples (Neg-distr), authors (Jin et al., 2020) enter limiting the global division of distance distributions (GDS) to two distribution to encourage clear separation of positive and negative samples from a global perspective. The paper (Lan et al., 2020) proposes a MagnifierNet with three branches, which precisely crushes parts from whole to parts, when it comes to distinguishing visually similar personalities or identifying a person with occlusion. The article (Fan et al., 2020) is presented RF-ReID, a novel approach that uses radio frequency (RF) signals for the long-term use of a person's ReID. These results also show two interesting features: First, because RF signals work in the presence of occlusion and poor lighting, RF-ReID allows human ReID to be used in such scenarios. Second, unlike photographs and videos that reveal personal and private information, RF signals are better at maintaining privacy. In article (Zhang et al., 2020a), authors propose an attentive feature aggregation module, namely, Multilevel Mindful Feature Aggregation Using Reference Data (MG-RAFA), to subtly aggregate spatio-temporal features into a distinctive feature representation at the video level. In article (Wang et al., 2020), a new framework by examining high-level relationships and topology information for distinctive features and robust reconciliation is proposed. The main idea is using the CNN backbone to explore feature maps and a key point scoring model to extract semantic local features. To improve the quality of pseudo-labels in existing methods, authors (Zeng et al., 2020) propose the HCT method, which combines hierarchical clustering with triplet loss in batch mode. The key idea behind HCT is to take full advantage of the similarity between samples in the target dataset through hierarchical clustering, reduce the impact of hard samples through hard packet loss of triplets, in order to generate high quality pseudolabels and improve model performance. In work (Zhang et al., 2020b), proposed an effective Relation-Aware Global Attention (RGA) module that captures global structural information for better attention training. In particular, for each position of the object, in order to compactly cover the structural information about the global scale and information about the local appearance, we propose to add the relations, that is, its pairwise correlations / similarities with all the positions of the objects (e. g., in raster scan order), and the function itself together to study attention using a shallow convolutional model. Authors (Kniaz et al., 2019) offer a ThermalGAN framework for cross-modal color thermal human re-identification (ReID). They use a set of generating adversarial networks (GANs) to transform a single color sensor image into a set of multimodal thermal sensors and thermal histograms and function descriptors as a thermal signature.

2.2 Generative Adversarial Networks

Generative adversarial networks (GANs) (Goodfellow et al., 2014) exploits an antagonistic game approach, that allows to significantly increase the quality of image-to-image translation (Isola et al., 2017, Zhang et al., 2017a, Zhang et al., 2017b). `pix2pix` GAN framework (Isola et al., 2017) carries out arbitrary image transformations, using geometrically aligned image pairs from source and target domains. The framework successfully performs arbitrary image-to-image translations such as season change and object transfiguration. The `pix2pix` network model (Zhang et al., 2017a, Zhang et al., 2017b) trained to transform a thermal image of a human face to the color image allows to improve the quality of a face recognition performance in a cross-modality thermal to visible range setting.

While human face has a relatively stable temperature, color-thermal image translation for more temperature-variable objects, such as the whole human body or vehicles with an arbitrary background, is more challenging.

3. METHOD

3.1 Framework Overview

Our aim is training two models that are jointly solve the person Re-ID task in the cross-modality setting. The first model is a `ThermalGAN` (Kniaz et al., 2019) generative network, that aims to translate a single color probe image into a multimodal set of thermal probe images. The second model is our `ParrotGAN` model that aim to train a latent space of person IDs. We consider three domains: an input color image domain $\mathcal{A} \in \mathbb{R}^{W \times H \times 3}$, a thermal image domain $\mathcal{B} \in \mathbb{R}^{W \times H}$, and the latent space domain $\mathcal{W} \in \mathbb{R}^{K}$, where W, H are image dimension, and K is the dimension of the latent space.

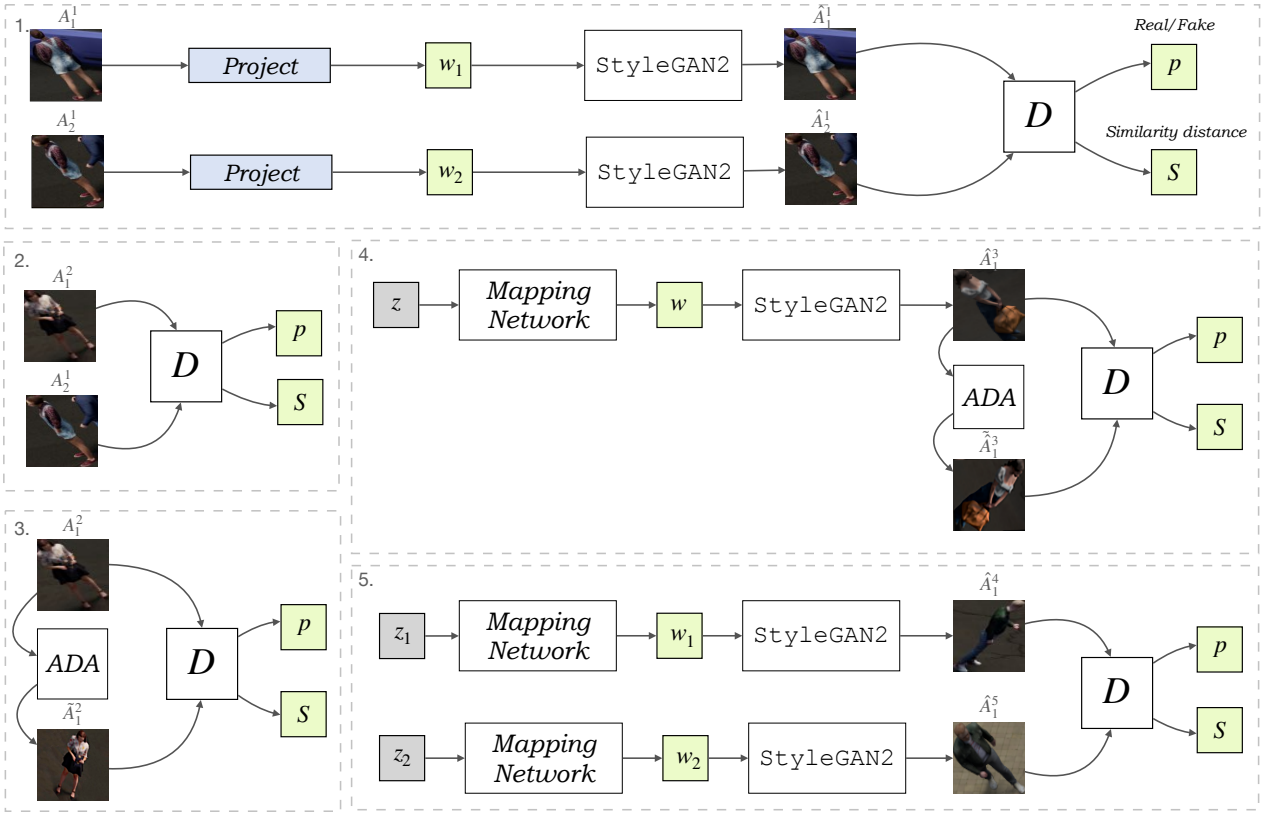

Figure 2. `ParrotGAN` **training process**: For each optimization iteration we perform five steps. (1) Back projection of real images of the same person ID to latent vectors. (2) Training discriminator on false pair. (3) Training discriminator on a true pair obtained using ADA augmentation of a real image. (4) Training discriminator on a true pair obtained using ADA augmentation of a fake image. (5) Training discriminator on a false pair generated using two random latent codes.

We use a `StyleGAN2` (Karras et al., 2020b) as a starting point for our `ParrotGAN` model. We modify the discriminator architecture. Our discriminator D receives a pair of person images $A_{j_1}^{i_1}, A_{j_1}^{i_2}$ as an input. We use upper index to represent the person ID $i \in \mathbb{N}$. The lower index denotes the index $j \in \mathbb{N}$ of the image in the probe or gallery set. The pair can present either the same person or two different persons. Also images in the pair can be either real of fake. The aim of our discriminator D is to predict two scalar values: a Wasserstein GAN (Gulrajani et al., 2017) penalty p, and a similarity distance d that must be equal zero for images of the same person, and equal one otherwise. We aim to train a mapping $D : (A_{j_1}^{i_1}, A_{j_1}^{i_2}) \rightarrow (p, s)$, where $p \in \mathbb{R}$ is a Wasserstein GAN penalty, and $s \in [0, 1]$ is a similarity distance.

Traditional GAN training algorithm includes two steps for each optimization iteration. Firstly, discriminator D is evaluated on real images. Secondly, fake images are synthesized by the generator G and evaluated by the discriminator D. In our `ParotGAN` framework, the discriminator D receives a pair of images that can be either real or fake, and represent the same or different person ID. Hence, we need more steps to train the discriminator using all possible input combinations.

3.2 `ParrotGAN`

We use five steps for each optimization iteration (Figure 2). At step (1), we project two real images of the same person ID A_1^1, A_2^1. We project images A_1^1, A_2^1 into the latent space of the `StyleGAN2` model to obtain latent codes w_1, w_2. After that we clear the gradients and synthesize fake images of the same person $\hat{A}_1^1, \hat{A}_2^1$. We evaluate the discriminator on the fake images requiring $s = 0$. At step (2), we evaluate the discriminator on two real images of different person IDs A_1^2, A_2^1. At step (3), we aim to evaluate discriminator on two real images of the same ID. We generate a new view $\tilde{A}_1^2$ of a real image A_1^2 using adaptive augmentation (Karras et al., 2020a). We evaluate the discriminator D using $A_1^2, \tilde{A}_1^2$ requiring $s = 1$. At step (4), we sample a random noise vector z and obtain a latent code w using the `StyleGAN2` mapping network. After that, we synthesize the fake image $\hat{A}_1^3$ and generate its novel view $\tilde{\hat{A}}_1^3$ using adaptive augmentation. We evaluate the discriminator D using $\hat{A}_1^3, \tilde{\hat{A}}_1^3$ requiring $s = 0$. Finally, at step (5), we sample two random noise vectors z_1, z_2 and generate two fake images of random person IDs $\hat{A}_1^3, \hat{A}_1^3$. We evaluate the discriminator D using these images requiring $s = 1$.

3.3 Loss functions

Three loss functions govern the training of our `ParrotGAN` model: the Wasserstein adversarial loss $\mathcal{L}_{adv}$, similarity loss $\mathcal{L}_s$, and the Re-ID loss $\mathcal{L}_d$. The aim of the Wasserstein adversarial loss $\mathcal{L}_{adv}$ is to distinguish real images A from the training dataset from the fake images $\hat{A}$ synthesized by the generator

$$\mathcal{L}_{adv} = \underset{\hat{A}\sim P_g}{E}[D(\hat{A})] - \underset{A\sim P_r}{E}[D(A)] + \lambda \underset{\hat{A}\sim P_A}{E}\left[\left(\left\|\nabla_{\hat{A}} D(\hat{A})\right\|_2 - 1\right)^2\right]. \quad (1)$$

The aim of our similarity loss $\mathcal{L}_s$ is to provide the gradient from the discriminator D forcing to learn a generator latent space, where the same person IDs are localized in compact subspaces

$$\mathcal{L}_s = \underset{A\sim P_r}{E}[s(A^m, A^n) \cdot \log \mathbb{1}_{m\neq n} - (1 - s(A^m, A^n)) \cdot \log(1 - \mathbb{1}_{m\neq n})], \quad (2)$$

where $s(A^m, A^n)$ is the similarity distance predicted by the discriminator D, $\mathbb{1}_{m\neq n}$ is an indicator function that is equal zero if A^m and A^n represent the same person ID, and is equal one otherwise. Our Re-ID loss $\mathcal{L}_d$ aims to provide disentanglement of the `StyleGAN2` latent space and inspired by the semi-supevised approach to disentanglement of a generative model (Nie et al., 2020)

$$\mathcal{L}_d = \underset{A\sim P_r}{E}\left[\left\|\mathbb{1}_{m=n} - \frac{P(A^m)\cdot P(A^n)}{\|A^m\|\,\|A^n\|}\right\|\right] \quad (3)$$

where $P : A \rightarrow w$ is a projection from an input image to the latent space. The final objective function $\mathcal{L}$ is obtained as the weighted sum of three losses

$$\mathcal{L} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_s\mathcal{L}_s + \lambda_d\mathcal{L}_d. \quad (4)$$

3.4 AerialReID dataset

We generated our *AerialReID* dataset using Blender 3D creation suite and high-quality 3D scans of real person IDs acquired using a structured light scanner (Knyaz, 2010). Our dataset includes full frame images from the virtual UAV with bounding box annotations of all humans, 30k pairs of cropped pairs of color and thermal images of 137 person IDs. Examples of full color and thermal images from our *AerialReID* dataset are given in Figure 3.

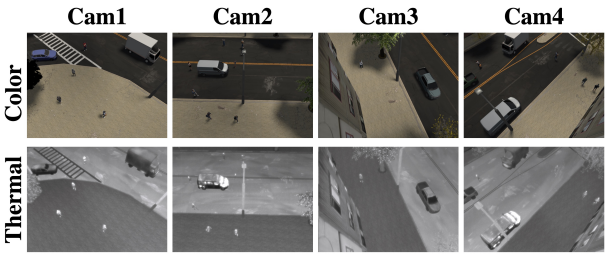

Figure 3. Examples of annotated images in our *AerialReID* dataset.

3D models were scanned from real humans and textured in visible and infrared ranges (Figure 4). Thermal textures were generated using the FLIR ONE PRO camera (Knyaz and Zheltov, 2018, Knyaz, 2019).

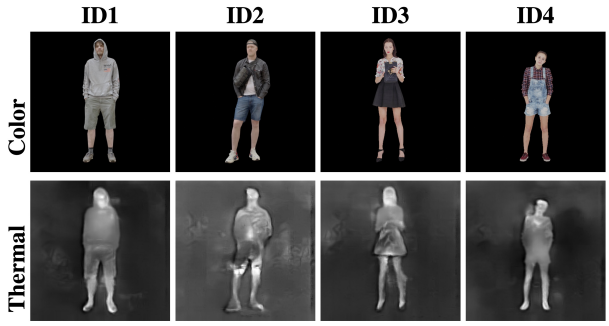

Figure 4. Examples of human 3D models textured in visible and infrared range in our *AerialReID* dataset.

We used a virtual 3D scene from the *SemanticVoxels* dataset (Kniaz et al., 2020, Kniaz and Moshkantseva, 2021). We placed ten cameras in our virtual environment. For each frame humans were placed in a random pose in the virtual 3D scene and random direction of a virtual light source was selected (Kniaz et al., 2021, Kniaz et al., 2023). Each frame was rendered in visible and thermal ranges at a 2k resolution corresponding to a typical modern UAV. Also the ground truth bounding boxes were generated for each frame. We cropped full frame images using the ground truth bounding boxes to obtain 30k images of person IDs (Figure 5).

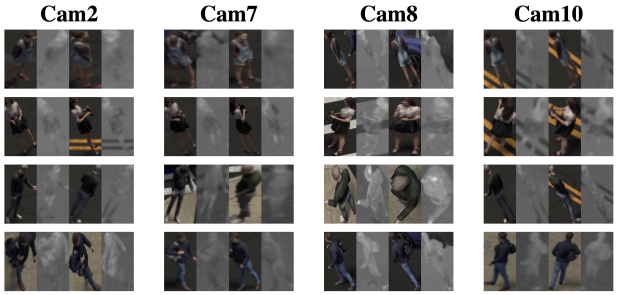

Figure 5. Examples of person images from our *AerialReID* dataset.

4. EXPERIMENTS

We evaluate our `ParrotGAN` framework and modern baselines using various metrics. We use the cumulative matching characteristic (CMC) curves and normalized area-under-curve (nAUC) for the ReID task. The evaluation demonstrated encouraging results and proved that our `ParrotGAN` framework outperforms existing baselines in the Re-ID accuracy. Furthermore, we demonstrated that the fusion of the semantic data with the input thermal gallery image increases the object detection and localization scores.

4.1 Network Training

We trained our models and baselines using train split of the *AerialReID* dataset. Training of the `ThremalGAN` model took 68 hours. We optimize network using minibatch SGD with an Adam solver. We use a learning rate of 0.0002, and momentum parameters $\beta_1 = 0.5$, $\beta_2 = 0.999$ similar to (Isola et al., 2017).

4.2 Quantitative Evaluation

We evaluate our model using our *AeriralReID* and *PRAI-1581* (Zhang et al., 2021) datasets. We evaluate our model and baselines quantitatively in terms of cumulative matching characteristic (CMC) curve and mean average precision (mAP). We compare our model to four model baselines. The `MuDeep` (Qian et al., 2017) model leverages two deep models. The results of the evaluation are presented in Table 1.

Table 1. Experiments on *AerialReID* dataset in single-shot setting.

Methods	ThermalWorld ReID single-shot					
	r = 1	r = 5	r = 10	r = 15	r = 20	mAP
MuDeep	15.2	25.6	34.4	38.1	40.0	20.9
ResNet-mid	9.1	15.4	26.5	30.1	35.3	16.1
HACNN	14.2	17.9	27.3	33.8	36.1	17.9
PCB	11.2	16.5	27.1	31.2	35.4	14.7
ParrotGAN	**17.3**	**29.1**	**36.6**	**39.9**	**42.3**	**22.6**

5. CONCLUSION

We demonstrated that the latent space of trained generative model can be effectively used for person Re-ID. Specifically, latent code obtained by back-projection of a person image into a latent space captures discriminative features that can be used to match corresponding person IDs. Furthermore, our model provides a robust performance in challenging scenarios such as person Re-ID in aerial images and cross-modality color-to-thermal person Re-ID.

REFERENCES

Ahmed, E., Jones, M., Marks, T. K., 2015. An improved deep learning architecture for person re-identification. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3908–3916.

Bak, S., Corvee, E., Bremond, F., Thonnat, M., 2010. Person re-identification using spatial covariance regions of human body parts. *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, IEEE, 435–440.

Bhuiyan, A., Perina, A., Murino, V., 2014. Person re-identification by discriminatively selecting parts and features. *European Conference on Computer Vision*, Springer, 147–161.

Bhuiyan, A., Perina, A., Murino, V., 2015. Person re-identification by discriminatively selecting parts and features. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Istituto Italiano di Tecnologia, Genoa, Italy, Springer International Publishing, Cham, 147–161.

Bhuiyan, A., Perina, A., Murino, V., 2018. Exploiting Multiple Detections for Person Re-Identification. *Journal of Imaging*, 4(2), 28.

Bird, N. D., Masoud, O., Papanikolopoulos, N. P., Isaacs, A., 2005. Detection of loitering individuals in public transportation areas. *IEEE Transactions on intelligent transportation systems*, 6(2), 167–177.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P., 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 2172–2180.

Cheng, D. S., Cristani, M., Stoppa, M., Bazzani, L., Murino, V., 2011. Custom pictorial structures for re-identification. *Bmvc*, 1, Citeseer, 6.

Fan, L., Li, T., Fang, R., Hristov, R., Yuan, Y., Katabi, D., 2020. Learning longterm representations for person re-identification using radio signals. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M., 2010a. Person re-identification by symmetry-driven accumulation of local features. *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2360–2367.

Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M., 2010b. Person re-identification by symmetry-driven accumulation of local features. *2010 IEEE computer society conference on computer vision and pattern recognition*, IEEE, 2360–2367.

GHEISSARI, N., 2006. Person Re-identification Using Spatiotemporal Appearance. *Proc. of CVPR, 2006*.

Gong, S., Cristani, M., Yan, S., 2014. *Person Re-Identification (Advances in Computer Vision and Pattern Recognition)*. Springer London, London.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. *Advances in neural information processing systems*, 2672–2680.

Gray, D., Tao, H., 2008. Viewpoint invariant pedestrian recognition with an ensemble of localized features. *European conference on computer vision*, Springer, 262–275.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A. C., 2017. Improved training of wasserstein gans. *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 5767–5777.

Hamdoun, O., Moutarde, F., Stanciulescu, B., Steux, B., 2008. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. *2008 Second ACM/IEEE International Conference on Distributed Smart Cameras*, IEEE, 1–6.

Isola, P., Zhu, J.-Y., Zhou, T., Efros, A. A., 2017. Image-to-Image Translation with Conditional Adversarial Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 5967–5976.

Jin, X., Lan, C., Zeng, W., Chen, Z., 2020. Global distance-distributions separation for unsupervised person re-identification. *European Conference on Computer Vision*, Springer, 735–751.

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T., 2020a. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T., 2020b. Analyzing and improving the image quality of stylegan. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 8107–8116.

Kniaz, V., Knyaz, V., Moshkantsev, P., 2023. Iq-gan: Instance-quantized image synthesis. B. Kryzhanovsky, W. Dunin-Barkowski, V. Redko, Y. Tiumentsev (eds), *Advances in Neural Computation, Machine Learning, and Cognitive Research VI*, Springer International Publishing, Cham, 277–291.

Kniaz, V. V., Knyaz, V. A., 2019. Chapter 6 - multispectral person re-identification using gan for color-to-thermal image translation. M. Y. Yang, B. Rosenhahn, V. Murino (eds), *Multimodal Scene Understanding*, Academic Press, 135–158.

Kniaz, V. V., Knyaz, V. A., Hladůvka, J., Kropatsch, W. G., Mizginov, V., 2019. Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset. L. Leal-Taixé, S. Roth (eds), *Computer Vision – ECCV 2018 Workshops*, Springer International Publishing, Cham, 606–624.

Kniaz, V. V., Knyaz, V. A., Mizginov, V. A., Papazyan, A., Fomin, N., Grodzitsky, L., 2021. Adversarial dataset augmentation using reinforcement learning and 3d modeling. B. Kryzhanovsky, W. Dunin-Barkowski, V. Redko, Y. Tiumentsev (eds), *Advances in Neural Computation, Machine Learning, and Cognitive Research IV. NEUROINFORMATICS 2020. Studies in Computational Intelligence*, 925, Springer International Publishing, Cham.

Kniaz, V. V., Knyaz, V. A., Remondino, F., Bordodymov, A., Moshkantsev, P., 2020. Image-to-voxel model translation for 3d scene reconstruction and segmentation. *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VII*, 105–124.

Kniaz, V. V., Moshkantseva, P., 2021. OBJECT RE-IDENTIFICATION USING MULTIMODAL AERIAL IMAGERY AND CONDITIONAL ADVERSARIAL NETWORKS. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIV-2/W1-2021, 131–136. https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLIV-2-W1-2021/131/2021/.

Knyaz, V., 2019. Multimodal data fusion for object recognition. *Proc. SPIE 11059, Multimodal Sensing: Technologies and Applications.*, 110590, 110590P. https://doi.org/10.1117/12.2526067.

Knyaz, V. A., 2010. Multi-media Projector – Single Camera Photogrammetric System For Fast 3d Reconstruction. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XXXVIII-5, 343–348. http://www.isprs.org/proceedings/XXXVIII/part5/papers/143.pdf.

Knyaz, V., Zheltov, S., 2018. Deep learning object recognition in multi-spectral UAV imagery. P. Schelkens, T. Ebrahimi, G. Cristóbal (eds), *Optics, Photonics, and Digital Technologies for Imaging Applications V*, 10679, International Society for Optics and Photonics, SPIE, 1067920.

Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P. M., Bischof, H., 2012. Large scale metric learning from equivalence constraints. *2012 IEEE conference on computer vision and pattern recognition*, IEEE, 2288–2295.

Lan, Y., Liu, Y., Tian, M., Zhou, X., Zhang, X., Yi, S., Li, H., 2020. MagnifierNet: Towards Semantic Adversary and Fusion for Person Re-identification. *arXiv preprint arXiv:2002.10979*.

Li, Z., Chang, S., Liang, F., Huang, T. S., Cao, L., Smith, J. R., 2013. Learning locally-adaptive decision functions for person verification. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3610–3617.

Mignon, A., Jurie, F., 2012. Pcca: A new approach for distance learning from sparse pairwise constraints. *2012 IEEE conference on computer vision and pattern recognition*, IEEE, 2666–2672.

Nguyen, D., Park, K., 2016. Body-Based Gender Recognition Using Images from Visible and Thermal Cameras. *Sensors*, 16(2), 156–21.

Nguyen, D. T., Hong, H. G., Kim, K. W., Park, K. R., 2017. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3), 605.

Nie, W., Karras, T., Garg, A., Debnath, S., Patney, A., Patel, A. B., Anandkumar, A., 2020. Semi-supervised stylegan for disentanglement learning. *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, 7360–7369.

Pedagadi, S., Orwell, J., Velastin, S., Boghossian, B., 2013. Local fisher discriminant analysis for pedestrian re-identification. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3318–3325.

Prosser, B., Gong, S., Xiang, T., 2008a. Multi-camera matching using bi-directional cumulative brightness transfer functions. *BMVC 2008 - Proceedings of the British Machine Vision Conference 2008*, Queen Mary, University of London, London, United Kingdom, British Machine Vision Association, 64.1–64.10.

Prosser, B. J., Gong, S., Xiang, T., 2008b. Multi-camera matching using bi-directional cumulative brightness transfer functions. *BMVC*, 8, Citeseer, 74.

Qian, X., Fu, Y., Jiang, Y., Xiang, T., Xue, X., 2017. Multi-scale deep learning architectures for person re-identification. *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, IEEE Computer Society, 5409–5418.

Quan, R., Dong, X., Wu, Y., Zhu, L., Yang, Y., 2019. Auto-reid: Searching for a part-aware convnet for person re-identification. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Schwartz, W. R., Davis, L. S., 2009. Learning discriminative appearance-based models using partial least squares. *2009 XXII Brazilian symposium on computer graphics and image processing*, IEEE, 322–329.

Tuzel, O., Porikli, F., Meer, P., 2008. Pedestrian detection via classification on riemannian manifolds. *IEEE transactions on pattern analysis and machine intelligence*, 30(10), 1713–1727.

Varior, R. R., Haloi, M., Wang, G., 2016a. Gated siamese convolutional neural network architecture for human re-identification. *European conference on computer vision*, Springer, 791–808.

Varior, R. R., Shuai, B., Lu, J., Xu, D., Wang, G., 2016b. A siamese long short-term memory architecture for human re-identification. *European conference on computer vision*, Springer, 135–153.

Wang, D., Zhang, S., 2020. Unsupervised person re-identification via multi-label classification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, G., Yang, S., Liu, H., Wang, Z., Yang, Y., Wang, S., Yu, G., Zhou, E., Sun, J., 2020. High-order information matters: Learning relation and topology for occluded person re-identification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, P., Jiao, B., Yang, L., Yang, Y., Zhang, S., Wei, W., Zhang, Y., 2019. Vehicle re-identification in aerial imagery: Dataset and approach. *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 460–469.

Wang, X., Doretto, G., Sebastian, T., Rittscher, J., Tu, P., 2007. Shape and appearance context modeling. *2007 ieee 11th international conference on computer vision*, Ieee, 1–8.

Wu, A., Zheng, W.-S., Yu, H.-X., Gong, S., Lai, J., 2017. RGB-Infrared Cross-Modality Person Re-Identification. *The IEEE International Conference on Computer Vision (ICCV)*.

Xiao, T., Li, H., Ouyang, W., Wang, X., 2016. Learning deep feature representations with domain guided dropout for person re-identification. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1249–1258.

Xiong, F., Gou, M., Camps, O., Sznaier, M., 2014. Person re-identification using kernel-based metric learning methods. *European conference on computer vision*, Springer, 1–16.

Yan, Y., Ni, B., Song, Z., Ma, C., Yan, Y., Yang, X., 2016. Person re-identification via recurrent feature aggregation. *European Conference on Computer Vision*, Springer, 701–716.

Ye, M., Lan, X., Li, J., Yuen, P. C., 2018a. Hierarchical Discriminative Learning for Visible Thermal Person Re-Identification. *AAAI*.

Ye, M., Wang, Z., Lan, X., Yuen, P. C., 2018b. Visible Thermal Person Re-Identification via Dual-Constrained Top-Ranking. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, International Joint Conferences on Artificial Intelligence Organization, California, 1092–1099.

Zeng, K., Ning, M., Wang, Y., Guo, Y., 2020. Hierarchical clustering with hard-batch triplet loss for person re-identification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, H., Patel, V. M., Riggan, B. S., Hu, S., 2017a. Generative adversarial network-based synthesis of visible faces from polarimetrie thermal faces. *2017 IEEE International Joint Conference on Biometrics (IJCB)*, IEEE, 100–107.

Zhang, S., Zhang, Q., Yang, Y., Wei, X., Wang, P., Jiao, B., Zhang, Y., 2021. Person Re-Identification in Aerial Imagery. *IEEE Trans. Multim.*, 23, 281–291. https://doi.org/10.1109/TMM.2020.2977528.

Zhang, T., Wiliem, A., Yang, S., Lovell, B. C., 2017b. TV-GAN: Generative Adversarial Network Based Thermal to Visible Face Recognition.

Zhang, Z., Lan, C., Zeng, W., Chen, Z., 2020a. Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, Z., Lan, C., Zeng, W., Jin, X., Chen, Z., 2020b. Relation-aware global attention for person re-identification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zheng, W.-S., Gong, S., Xiang, T., 2011. Person re-identification by probabilistic relative distance comparison. *CVPR 2011*, IEEE, 649–656.

Zheng, W.-S., Gong, S., Xiang, T., 2012. Reidentification by relative distance comparison. *IEEE transactions on pattern analysis and machine intelligence*, 35(3), 653–668.