

# Variational Autoencoders for Pedestrian Synthetic Data Augmentation of Existing Datasets: A Preliminary Investigation

Ivan Nikolov <sup>a</sup>

*Computer Graphics Group, Department of Architecture, Design and Media Technology,  
Aalborg University, Aalborg, Denmark*

**Keywords:** Synthetic Data, Variational Autoencoders, Object Detection, Dataset Augmentation, Surveillance.

**Abstract:** The requirements for more and more data for training deep learning surveillance and object detection models have resulted in slower deployment and more costs connected to dataset gathering, annotation, and testing. One way to help with this is the use of synthetic data giving more varied scenarios and not requiring manual annotation. We present our initial exploratory work in generating synthetic pedestrian augmentations for an existing dataset through the use of variational autoencoders. Our method consists of creating a large number of backgrounds and training a variational autoencoder on a small subset of annotated pedestrians. We then interpolate the latent space of the autoencoder to generate variations of these pedestrians, calculate their positions on the backgrounds, and blend them to create new images. We show that even though we do not achieve as good results as just adding more real images, we can boost the performance and robustness of a YoloV5 model trained on a mix of real and small amounts of synthetic images. As part of this paper, we also propose the next steps to expand this approach and make it much more useful for a wider array of datasets.

## 1 INTRODUCTION


Deep learning models require larger and larger datasets, which in many cases require specialized data not available freely online. This is especially true for vision models deployed for specific use cases like autonomous surveillance of indoor or outdoor scenes, anomaly detection, traffic analysis, etc. In these cases, if there is no easy access to comparable open-source datasets, the training and testing data needs to be captured for long periods, manually annotated, and then used for training and verification of the specific model. This can significantly slow down deployment and can put a large monetary and development burden upfront (de Melo et al., 2022). In many cases even if useful datasets exist online, they may not contain the required images or videos in large enough quantities to provide sufficient resources for training robust enough models based on modern architectures like transformers (Nikolenko, 2021).

Another widely spread problem with gathering datasets is that in some scenarios installing and configuring the required hardware and capturing footage can be a dangerous or impossible process that could require trained specialists and necessary permits.

Looking from an ethical standpoint capturing data can also infringe on people's privacy, even if it is done in outdoor scenes (Voigt and Von dem Bussche, 2017).

As mentioned before capturing the necessary data is only the start of the process, after which it needs to be annotated, which can pose an even larger burden. Even with modern labeling tools like Labelbox, CVAT, Labelerr, among others, and the emergence of companies focused on providing services for the annotation of data by volunteers like Humans in the Loop (in the Loop, 2018; V7Labs, 2019), the creation of fine-grained bounding boxes and pixel annotations can take a lot of time and money.

This scarcity of image data has prompted researchers to look more and more into synthetically generating data for deep learning. We can separate the generation of synthetic data into two main types - based on the use of digital twins and based on deep learning models. Research based on digital twins focuses on the modeling of real-world environments (Ros et al., 2016; Wang et al., 2019), objects (Nikolov, 2023; Acsintoae et al., 2022), and phenomena (Halder et al., 2019) and using them to reproduce synthetic approximations of the scene from which images and videos can be captured. Deep learning-generated synthetic data relies mostly on large generative models

<sup>a</sup>  <https://orcid.org/0000-0002-4952-8848>

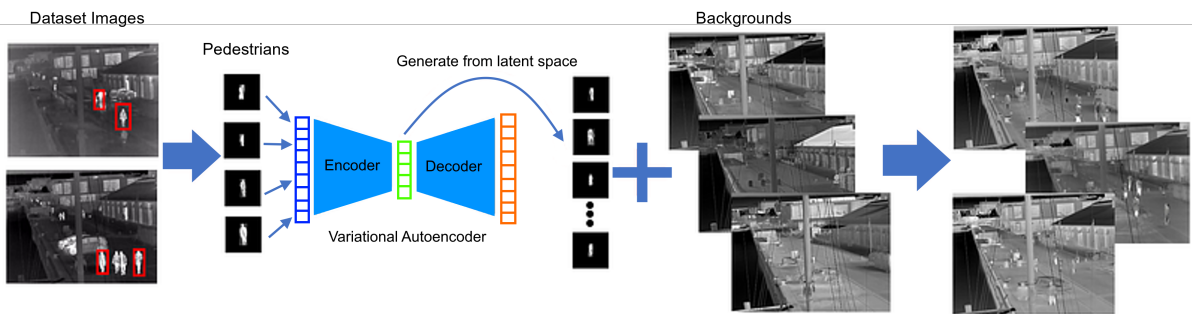


Figure 1: Overview of the proposed generation of synthetic pedestrians using a VAE for image augmentation.

like Dall-E, Stable Diffusion, and Midjourney (Abduljawad and Alsalmami, 2022; Borji, 2022), as well as Generative adversarial networks (GANs) and autoencoders (Huang et al., 2018; Brock et al., 2018; He et al., 2022; Islam and Zhang, 2020). Many of these approaches focus on generating or manipulating the full image from a given prompt or a starter image, which require much larger models and are susceptible to distribution gaps between synthesized and real images. On the other hand augmenting parts of images with synthetic elements a lot of times requires composing and blending with semi-realistic objects (Chan et al., 2021; Tripathi et al., 2019).

We propose a preliminary study in the generation of synthetic pedestrian variations for augmenting surveillance images. This way we can easily extend existing datasets with more pedestrian visualizations, thus making them more varied and helping train more robust deep learning algorithms on them. We choose to use the LTD (Nikolov et al., 2021) dataset as part of this paper, but the proposed pipeline can be easily extended to other datasets. Our approach consists of training a deep feature consistent variational autoencoder (VAE) (Hou et al., 2017) on a small subset of pedestrian images and sampling its latent space to generate interpolations of pedestrians. We then compose and blend these interpolations using the Poisson Image Editing (Pérez et al., 2023) method together with the extracted background images from the dataset to generate synthetic variations with different numbers of people at different positions.

We do an initial testing of our augmented data by training a YoloV5 (Jocher et al., 2020) model and comparing its performance on detecting pedestrians when trained only on real data, several mixed variations between real and augmented synthetic data and only augmented synthetic data. We show that the mixed datasets where only a small amount of synthetic data is given have the potential to boost the performance and robustness of the models against data drift.

The contributions of this early-stage research are:

1. We propose a lightweight and easily transferable approach for generating synthetic variations for augmenting datasets with accompanying ready-made bounding box annotations, that require relatively little manual preparation work;
2. We do an initial test of the feasibility of the approach for training models with less real data and in some cases boosting their performance;
3. We propose the next steps to expanding this approach, making it more robust and more useful for a wider range of datasets.

In the next section, we will give an overview of the related work on surveillance datasets and generating synthetic data. In Section 3 we will present the overview of our approach and its different components. Finally, in Section 4 we will show our initial results together with a proposal for the next steps in expanding the pipeline.

## 2 RELATED WORK

In this section, we will focus on both existing real-world surveillance and anomaly detection datasets, as well as the different approaches to generating synthetic datasets.

### 2.1 Surveillance Datasets

Surveillance datasets can be roughly separated into two categories (Nikolov et al., 2021). The first one focuses on changing backgrounds representing images and videos captured from movie perspectives like vehicles and egocentric views. The second category is directed towards stationary views captured from surveillance CCTV cameras.

From stationary background datasets, most are directed towards anomaly detection and pedestrian detection and tracking use cases. These datasets are captured from cameras sitting in one place for varying

amounts of time. Most of these datasets also are captured for very short amounts of time giving them a limited variation of scenarios, pedestrians, and vehicles (Lu et al., 2013; Liu et al., 2018). The exception to this are the MEVA (Corona et al., 2021) and LTD (Nikolov et al., 2021) datasets, which focus on larger periods and diverse environmental conditions and pedestrians. Even with the larger datasets, the problem persists that they are finite and do not provide all possible pedestrian scenarios to train robust enough models. In addition to this, annotating these datasets can be challenging and time-consuming.

## 2.2 Synthetic Data Generation

Synthetic datasets can be divided mainly into two groups - ones that are fully generated from scratch either by a deep learning model (Borji, 2022; He et al., 2022) or through digital twin technology (Wang et al., 2019; Grcić et al., 2021) and others that rely on augmenting existing datasets with synthetic parts and objects depending on the use case. Fully synthetically generated datasets oftentimes have the problem of a larger distribution gap between the synthetic and real-world data and require additional post-processing before training (Acscintoae et al., 2022). Datasets that only contain augmented synthetic elements, often fare better as they contain real-world backgrounds and objects together with the synthetic ones. In both cases, synthesizing data has the potential to create hard-to-reproduce scenarios, small data variations or emergencies, and anomalies that would be hard to capture in real life (Nikolenko, 2021).

The LTD dataset is one of the largest datasets for surveillance. It contains different weather conditions and times of day, together with variations in pedestrian activity, vehicles, bicycles, and even ships. Together with its newer extension providing annotations for large parts of the dataset, this makes it an ideal candidate for testing synthetic data generation methodologies, as shown from the work of Madan et. al (Madan et al., 2023). We chose this dataset to demonstrate our proposed solution as it has been proven to provide a challenging environment for out-of-distribution testing data, as shown from the results of anomaly and pedestrian detection models given in the dataset paper presenting it. Thus we can show if our proposed synthetic data can alleviate these problems.

## 3 PROPOSED METHOD

We propose the initial exploration of a method for generating synthetic pedestrian variations by exploring the latent space of a variational autoencoder trained on a small subset of pedestrian images. This section presents the different parts of the generation and augmentation process. The full pipeline is shown in Figure 1.

We first extract real background images from the LTD dataset, the same way it is presented by Madan et. al (Madan et al., 2023) We then select a subset of pedestrian annotations from the ground truth of the LTD dataset and pre-process them so they can be used as inputs for the selected VAE model. We train the selected model on the pre-processed data and extract from its latent space linear interpolations between two given pedestrian inputs. We then select a random interpolated output from the VAE and blend it into the background on a position that is also calculated as an interpolation from the given pedestrians' positions. For blending we use the Poisson Image Editing (Pérez et al., 2023) method for smooth blending. In this way, we can augment different number of pedestrians into each background image. The pipeline of our proposed solution is explained in the new subsections as follows - 3.1 Background Generation, 3.2 Pedestrian Extraction and Pre-processing, 3.3 VAE model Explanation, and 3.4 Generation of Pedestrians and Blending.

### 3.1 Background Generation

For generating the background images used for the augmentation process we use the process proposed by Madan et. al (Madan et al., 2023). As we currently generate only still images, we use a temporal median filter on the full image, without a specified mask. As the dataset spans 8 months of day and night footage and we want to create widely varying backgrounds we uniformly sample one background from the captured 2-minute videos once every two hours. We do this process for one week out of each of the 8 months, which gives us 672 background images. Later on, we will select at random from these backgrounds when blending them with the generated pedestrian images.

### 3.2 Pedestrian Extraction and Pre-Processing

To gather enough data to train the VAE model we use the small annotated training subsets given as part of the LTD dataset (Nikolov et al., 2021) - for the daily, weekly, and monthly February data. Even though

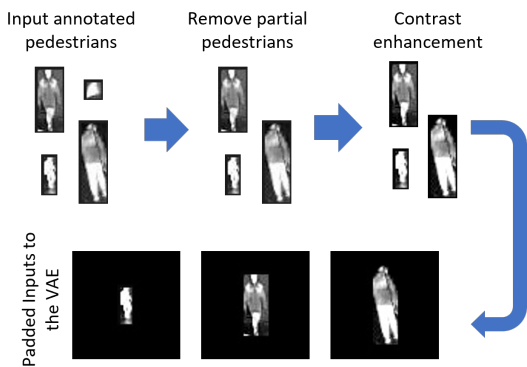


Figure 2: Pre-processing of the extracted pedestrian sub-images before using them as input for the VAE.

there are more annotations in the second iteration of the dataset we wanted to show the performance of our proposed method when data is generated from a small subset of annotations, when researchers do not have the means or time to annotate more. We use the bounding box annotations to extract the pedestrian sub-images. We filter the extracted sub-images to remove those that do not have heights at least twice as large as the width. This is done to remove annotations of pedestrians behind vehicles, parts of the background, or other pedestrians, which would be out of place if augmented into the background images. As the chosen VAE tends to blur the images and the inputs are very low-resolution, we also use a sharpening filter on them to pronounce the smaller details. We then reshape all pedestrian images to the same size by adding black borders around the sub-images to get it to a size of 128x128 pixels. This way the autoencoder can be fed the same size inputs without distorting the sub-images. The pedestrian pre-processing pipeline is given in Figure 2.

### 3.3 VAE Model Explanation

To generate the pedestrian variations we use a deep feature consistent variational autoencoder (Hou et al., 2017). Even though there are much more complex autoencoders like VQ-VAE2 (Razavi et al., 2019), DIP VAE (Kumar et al., 2017), MIWAE (Rainforth et al., 2018), which can provide better reconstructions, but require more data and more resources to train, we wanted to test our initial idea with a relatively straightforward architecture. We train the VAE using 1000 pre-processed pedestrian images. The training is done for 100 epochs, with a batch size of 8, using the combination of binary cross entropy and Kullback–Leibler divergence loss (Kingma and Welling, 2013). The latent space of the VAE was set to 1000 and the input image size was set to 128x128.

We have chosen the size of the latent space to ensure a large enough information is learned by the autoencoder to be able to represent the input images and interpolate them. In the paper proposing the model, the authors use a latent space of only 100, which is large enough for capturing the use case of human faces they present. After heuristically exploring the results from using different sizes of latent space we come to the conclusion that a size of 1000 helps the model learn the smaller details of the pedestrians, as for some of them only a couple of pixels can represent body parts. Examples of inputs together with interpolated latent space outputs are shown in Figure 3.

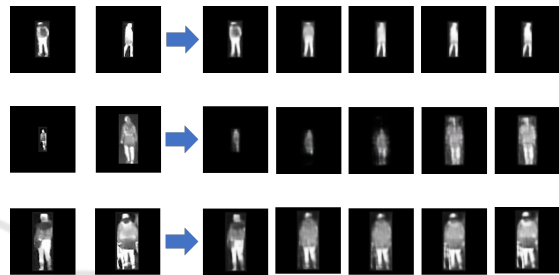


Figure 3: Example pedestrian inputs to the VAE (left two) and outputs from interpolating the latent space (right ones).

### 3.4 Generation of Pedestrians and Blending

Once we have the backgrounds created and the VAE trained, we create a workflow that first selects a background image randomly. Then two of the input pedestrians are selected at random from the same training set and run through the encoder part of the VAE. The resultant latent space encodings between the two images are interpolated with the desired number of steps. We use the same linear interpolation of the space, presented in the paper proposing the VAE (Hou et al., 2017), where the interpolation is defined by a linear transformation  $z = (1 - \alpha) * z_{left} + \alpha * z_{right}$ , where  $z_{left}$  and  $z_{right}$  are the latent vectors of the two input pedestrian images and  $\alpha = 0, 0.1, 0.2, \dots, 1$ . Heuristically we have seen that between 10 and 15 steps give enough variation with noticeable differences between the images created from the interpolations. These interpolations are then run through the decoder part of the VAE to produce images that are interpolations between the two inputs. We have selected to use a VAE for this, instead of a simple linear interpolation between the images as extracting interpolation from the latent space would result in more meaningful and gradual changes between the images and more visually coherent and pleasing results. It also minimizes

noticeable visual artifacts and is more likely to generate images that contain the same overall characteristics as the input images. This is especially important as we are working with very low-resolution pedestrian images, where each pixel captures a lot of information about the object.

We then select at random one of the generated images from the interpolation sequence and use linear interpolation between the coordinates of the two input sub-images to calculate its position in the larger background image. Finally, we blend between the background and the generated pedestrian image using the Poisson Image Editing (Pérez et al., 2023) algorithm. This algorithm removes many of the border artifacts between the pedestrian image and the background but also has the added benefit of changing the pedestrian’s color to better match the background. This is especially important as the background is taken from different months and times of day compared to the pedestrian images. This process is performed as many times as we need synthetic pedestrian augmentations on the background image. After observing the distributions of pedestrians in the real images we select that between 1 and 15 pedestrians are augmented in the proposed way for each synthetic image. A comparison between real images from the LTD dataset and synthetic images from our proposed solution is given in Figure 4.

We can see that even after the blending some of the synthetic pedestrian images contain darker parts and visual artifacts around the edges when the difference between the background and pedestrians is too big. We will discuss ideas on how to mitigate this in Section 5.

## 4 EXPERIMENTS AND RESULTS

To test out the proposed synthetic data augmentation and the generated data, we choose to train a YoloV5 model to detect pedestrians. We have seen from the work connected to the LTD paper, that the YoloV5 model had problems when tested on data from other months than the one it was trained on. We propose to see if training the YoloV5 model on a combination of real and synthetically augmented data can help the model perform better with more diverse data, effectively minimizing the data drift.

To this extent, we select six training datasets. The first two datasets are comprised only of real data present in the LTD data. The dataset contains thermal videos of size 288x384, captured with a Hikvision DS-2TD2235D-25/50 thermal camera. The clips are 8-bit grayscale. The videos are separated into

frames. First, the February month subset contains 200 images ( $D_{r200}$ ) captured from the full month both through the day and night. The second subset is a combination of the February month and March week datasets comprising 300 images ( $D_{r300}$ ). This was chosen to see how much performance is gotten from training the model on more real data. The other three datasets contain synthesized images together with the real ones. We wanted to see how the presence of synthetic data would influence the performance of the YoloV5 model, so the datasets contain the real February month 200 images, plus 100 ( $D_{m300}$ ), 200 ( $D_{m400}$ ), and 800 ( $D_{m1000}$ ) synthetic training images. With this, we wanted to see how the algorithm performs when the synthetic data is less than the real one, when it is equal to it, and when it is the larger part. Finally, we also have one dataset comprised of only 3000 synthetic images ( $D_{s3000}$ ).

We use each of the six datasets to train the YoloV5s pre-trained model, with a batch size of 16 for 100 epochs. We use the validation dataset provided as part of the LTD dataset and test on the three test datasets for the January, April, and August months. Each of the testing datasets contains 100 images and annotations.

For evaluation metrics, we calculate the precision, recall, as well as the mean average precision at an intersection over union (IoU) threshold 0.50 -  $mAP_{50}$  and at a varying IoU threshold between 0.50 and 0.95  $mAP_{50-95}$ . The results are given in Table 1.

We can see from the table that adding more images from another month ( $D_{r300}$ ) helps with the performance of the model especially for the April and August testing subsets. Using additional synthetic data in  $D_{m300}$  also results in a model that is more robust towards data drift, even if the performance boost is not as strong as adding additional real data. Adding more synthetic data with  $D_{m400}$  and  $D_{m1000}$  does not result in better results but is seen to degrade performance in most cases, except for the  $mAP_{50}$  for April in the case of using 800 additional synthetic images. When trained only on synthetic data the model does not learn enough to be useful for detecting real pedestrians. This shows that even though the initial results are positive in boosting model performance without the need for the annotation of additional real data, there is still a distribution gap between the real and synthetic pedestrians. This is most probably caused by the blurry outputs from the simpler VAE and the imperfect blending in some cases. We will discuss possible ways to address these problems and extend the proposed approach.

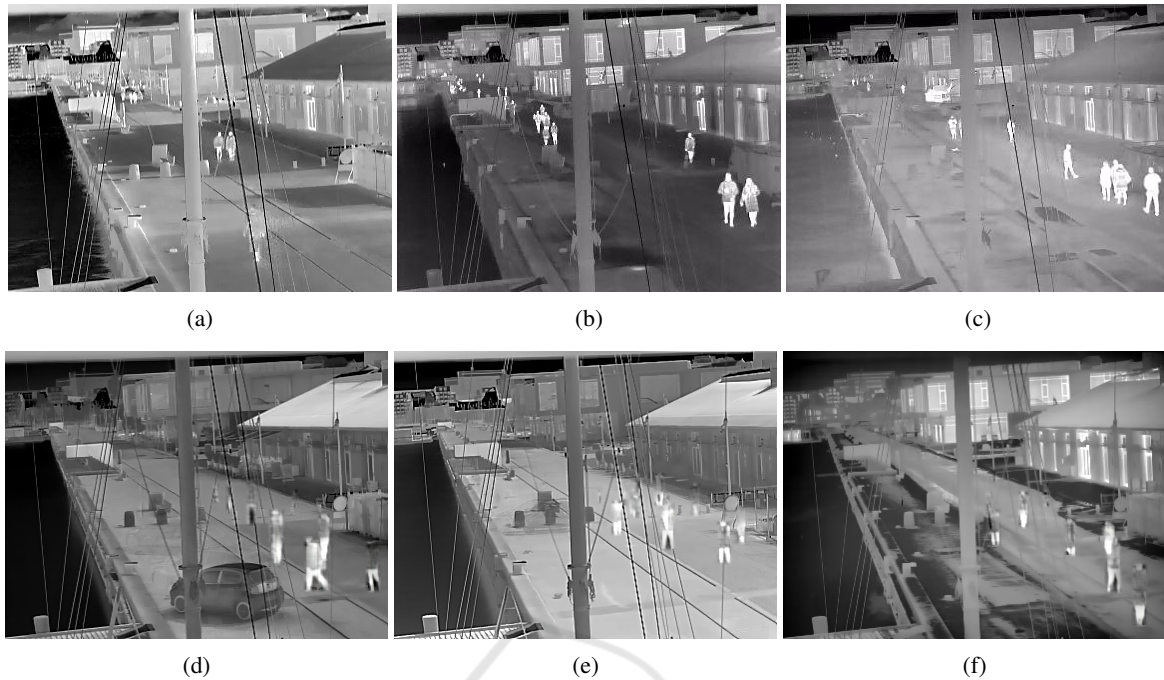


Figure 4: Visual comparison between the real images taken from the LTD dataset (top images) and the augmented synthetic images using the VAE latent space interpolation outputs (bottom images).

Table 1: Results from training YOLOv5 on the 6 chosen subsets of data - February month 200 real images ( $D_{r200}$ ), February month plus March week 300 real images ( $D_{r300}$ ), February month 200 real images plus 100 ( $D_{m300}$ ), 200 ( $D_{m400}$ ), and 800 ( $D_{m1000}$ ) synthetic images and 3000 synthetic images only ( $D_{s3000}$ ). The models are then tested on the January, April, and August test data from the LTD dataset.

Datasets	January Test Data		April Test Data		August Test Data	
	$mAP_{50}$	$mAP_{50-95}$	$mAP_{50}$	$mAP_{50-95}$	$mAP_{50}$	$mAP_{50-95}$
$D_{r200}$	0.809	0.461	0.476	0.213	0.512	0.243
$D_{r300}$	0.826	0.492	0.620	0.301	0.585	0.282
$D_{m300}$	0.81	0.458	0.524	0.259	0.548	0.246
$D_{m400}$	0.795	0.415	0.494	0.194	0.494	0.222
$D_{m1000}$	0.793	0.420	0.530	0.222	0.487	0.215
$D_{s3000}$	0.397	0.171	0.238	0.079	0.244	0.091

## 5 NEXT STEPS

The next steps for developing the proposed synthetic data generation algorithm would be to swap the simple VAE with some of the newer more complex and robust models like VQ-VAE2, DIP VAE, MIWAE, etc. This is especially evident from looking at the generated pedestrian variations, which have a great deal of blurring and artifacts. This is especially problematic as the pedestrian images are very low-resolution and any loss of details can be problematic for generating useful data. Additionally, the more complex variational autoencoders would give us the possibility to generate data for higher-resolution datasets like

Avenue (Lu et al., 2013) and ShanghaiTech Campus (Liu et al., 2018), in which the pedestrians have much more detail.

In the current research, we show that even with the simple VAE encoder augmentations, the additional synthetic images helped the YOLO model be more robust to the visual changes in the training data over time. We would like to do two additional studies to further prove that the proposed solution is the thing that improves performance. First, we would like to compare augmentations created through variational autoencoders with more straightforward approaches like linear interpolation between two pedestrian images and not through the linear space created by the

VAE. Second, an ablation study that removes parts of the proposed pipeline like Poisson Image Editing, contrast enhancement, etc. to see which has the most effect on the performance of the pipeline.

Another problem that can be addressed is smoothing the visual borders between the generated pedestrian's background and the background image. Currently even after the blending algorithm in some cases when the background is much brighter or darker than the pedestrian's background, some artifacts remain. One way to prevent that from happening is to use a segmentation model like ClipSEG (Lüddecke and Ecker, 2022) on the generated synthetic image and only capture the pixels that belong to the pedestrian or even parts of their bodies. This way the borders will be less of a problem.

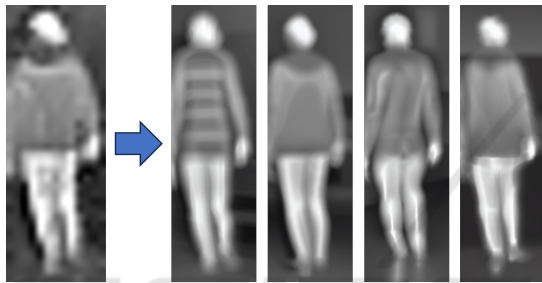


Figure 5: Example of possible Stable Diffusion variation outputs that can be augmented into the background.

Finally, to expand this proposal we would like to combine the synthetic pedestrians with a generative model like Stable Diffusion (Rombach et al., 2022), which can additionally augment them and lower the chance of the synthetic data being less varied and overfitting models trained on it. An example of such augmentation of the synthetic pedestrians is given in Figure 5.

## 6 CONCLUSION

We presented our initial exploration into creating synthetic pedestrian data augmentation for surveillance tasks using variational autoencoders. Synthetic data has become more and more widely used in deep learning tasks, especially in anomaly and object detection. We propose a lightweight approach to generate synthetic augmentation for existing datasets by blending interpolated variations from the latent space of a VAE trained on pedestrian data into pre-made background images extracted from the dataset.

To test our proposed pipeline we train a YOLOv5 object detector on real, synthetic, and mixed data from the LTD dataset. We show that even though the

synthetic data does not result in better performance than adding more real data, we see a performance uptick with testing data that is farther from the training one. This shows that the introduction of synthetic data can make the model more robust to data drift. The generated pedestrians are too simplistic and without enough variation, which results larger distribution gap between real and synthetic data. We propose ways to alleviate these problems by employing newer variational autoencoders and using segmentation models to better separate the models from the background and make blending easier. Even with the imperfect results from this initial research, we show that there is usefulness in generating synthetic humans through the use of exploring the latent space of VAEs.

## REFERENCES

- Abduljawad, M. and Alsalmami, A. (2022). Towards creating exotic remote sensing datasets using image generating ai. In *2022 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, pages 84–88. IEEE.
- Acsintoae, A., Florescu, A., Georgescu, M.-I., Mare, T., Sumedrea, P., Ionescu, R. T., Khan, F. S., and Shah, M. (2022). Unnormal: New benchmark for supervised open-set video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20143–20153.
- Borji, A. (2022). Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2. *arXiv preprint arXiv:2210.00586*.
- Brock, A., Donahue, J., and Simonyan, K. (2018). Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.
- Chan, R., Lis, K., Uhlemeyer, S., Blum, H., Honari, S., Siegwart, R., Fua, P., Salzmann, M., and Rottmann, M. (2021). Segmentmeifyoucan: A benchmark for anomaly segmentation. *arXiv preprint arXiv:2104.14812*.
- Corona, K., Osterdahl, K., Collins, R., and Hoogs, A. (2021). Meva: A large-scale multiview, multimodal video dataset for activity detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1060–1068.
- de Melo, C. M., Torralba, A., Guibas, L., DiCarlo, J., Chellappa, R., and Hodgins, J. (2022). Next-generation deep learning based on simulators and synthetic data. *Trends in cognitive sciences*.
- Grcić, M., Bevandić, P., and Šegvić, S. (2021). Dense anomaly detection by robust learning on synthetic negative data. *arXiv preprint arXiv:2112.12833*.
- Halder, S. S., Lalonde, J.-F., and Charette, R. d. (2019). Physics-based rendering for improving robustness to rain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10203–10212.

- He, R., Sun, S., Yu, X., Xue, C., Zhang, W., Torr, P., Bai, S., and Qi, X. (2022). Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*.
- Hou, X., Shen, L., Sun, K., and Qiu, G. (2017). Deep feature consistent variational autoencoder. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 1133–1141. IEEE.
- Huang, H., He, R., Sun, Z., Tan, T., et al. (2018). Introvae: Introspective variational autoencoders for photographic image synthesis. *Advances in neural information processing systems*, 31.
- in the Loop, H. (2018). Humans in the loop. <https://humansintheloop.org/>. Accessed: 2023-11-22.
- Islam, J. and Zhang, Y. (2020). Gan-based synthetic brain pet image generation. *Brain informatics*, 7:1–12.
- Jocher, G., Stoken, A., Borovec, J., NanoCode012, ChristopherSTAN, Changyu, L., Laughing, tkianai, Hogan, A., lorenzomamma, yxNONG, AlexWang1900, Diaconu, L., Marc, wanghaoyang0106, ml5ah, Doug, Ingham, F., Frederik, Guilhen, Hatovix, Poznanski, J., Fang, J., Yu, L., changyu98, Wang, M., Gupta, N., Akhtar, O., PetrDvoracek, and Rai, P. (2020). ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements.
- Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *CoRR*, abs/1312.6114.
- Kumar, A., Sattigeri, P., and Balakrishnan, A. (2017). Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*.
- Liu, W., W. Luo, D. L., and Gao, S. (2018). Future frame prediction for anomaly detection – a new baseline. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Lu, C., Shi, J., and Jia, J. (2013). Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727.
- Lüddecke, T. and Ecker, A. (2022). Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7086–7096.
- Madan, N., Siemon, M. S. N., Gjerde, M. K., Petersson, B. S., Grotuzas, A., Esbensen, M. A., Nikolov, I. A., Philipsen, M. P., Nasrollahi, K., and Moeslund, T. B. (2023). Thermalsynth: A novel approach for generating synthetic thermal human scenarios. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 130–139.
- Nikolenko, S. I. (2021). *Synthetic Data for Deep Learning*. Number 978-3-030-75178-4 in Springer Optimization and Its Applications. Springer.
- Nikolov, I. (2023). Augmenting Anomaly Detection Datasets with Reactive Synthetic Elements. In Vangorp, P. and Hunter, D., editors, *Computer Graphics and Visual Computing (CGVC)*. The Eurographics Association.
- Nikolov, I. A., Philipsen, M. P., Liu, J., Dueholm, J. V., Johansen, A. S., Nasrollahi, K., and Moeslund, T. B. (2021). Seasons in drift: A long-term thermal imaging dataset for studying concept drift. In *Thirty-fifth Conference on Neural Information Processing Systems*. Neural Information Processing Systems Foundation.
- Pérez, P., Gangnet, M., and Blake, A. (2023). Poisson image editing. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 577–582.
- Rainforth, T., Kosiorek, A., Le, T. A., Maddison, C., Igl, M., Wood, F., and Teh, Y. W. (2018). Tighter variational bounds are not necessarily better. In *International Conference on Machine Learning*, pages 4277–4285. PMLR.
- Razavi, A., Van den Oord, A., and Vinyals, O. (2019). Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.
- Ros, G., Sellart, L., Materzynska, J., Vazquez, D., and Lopez, A. M. (2016). The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243.
- Tripathi, S., Chandra, S., Agrawal, A., Tyagi, A., Rehg, J. M., and Chari, V. (2019). Learning to generate synthetic data via compositing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 461–470.
- V7Labs (2019). V7 labs annotation. <https://www.v7labs.com/>. Accessed: 2023-11-22.
- Voigt, P. and Von dem Bussche, A. (2017). The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed.*, Cham: Springer International Publishing, 10(3152676):10–5555.
- Wang, Q., Gao, J., Lin, W., and Yuan, Y. (2019). Learning from synthetic data for crowd counting in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8198–8207.