Using Synthetic Data for Person Tracking Under Adverse Weather Conditions

Abdulrahman Kerim[a,c,*], Ufuk Celikcan[a,*], Erkut Erdem[a], Aykut Erdem[b]

[a]*Hacettepe University, Department of Computer Engineering, Ankara, Turkey*
[b]*Koç University, Department of Computer Engineering, Istanbul, Turkey*
[c]*Lancaster University, School of Computing and Communications, UK*

Abstract

Robust visual tracking plays a vital role in many areas such as autonomous cars, surveillance and robotics. Recent trackers were shown to achieve adequate results under normal tracking scenarios with clear weather conditions, standard camera setups and lighting conditions. Yet, the performance of these trackers, whether they are correlation filter-based or learning-based, degrade under adverse weather conditions. The lack of videos with such weather conditions, in the available visual object tracking datasets, is the prime issue behind the low performance of the learning-based tracking algorithms. In this work, we provide a new person tracking dataset of real-world sequences (PTAW172Real) captured under foggy, rainy and snowy weather conditions to assess the performance of the current trackers. We also introduce a novel person tracking dataset of synthetic sequences (PTAW217Synth) procedurally generated by our NOVA framework spanning the same weather conditions in varying severity to mitigate the problem of data scarcity. Our experimental results demonstrate that the performances of the state-of-the-art deep trackers under adverse weather conditions can be boosted when the available real training sequences are complemented with our synthetically generated dataset during training.

Keywords:
Person Tracking, Synthetic Data, Rendering, Procedural Generation

[*] Corresponding authors
Email addresses: `a.kerim@lancaster.ac.uk` (Abdulrahman Kerim),

1. Introduction

Recently, convolutional neural networks (CNN) have shown a remarkable progress in various computer vision tasks such as object detection [1], object tracking [2], semantic segmentation [3], depth estimation [4], optical flow estimation [5], and person re-identification (ReID) [6]. Compared to the traditional shallow approaches, CNN-based models exhibit better generalization ability and perform much more accurately. Moreover, their performance usually increases as their expressive power increases, i.e. having more layers and/or more parameters. Yet, this introduces another difficulty as training such big networks requires more data and more powerful computing devices. The introduction of cheap general purpose graphics processing units (GPGPUs) have alleviated the hardware limitations. However, the scarcity of large-scale datasets for training supervised learning methods still remains as the main bottleneck for many computer vision tasks, especially, the ones that require enormous efforts for annotation, such as semantic segmentation and visual object tracking. Besides, for some others, e.g. optical flow and depth estimation, it becomes virtually impossible to provide large-scale densely annotated datasets.

In addition to the aforementioned need for large-scale benchmark datasets, another requirement is to have a high level of diversity to allow deep learning models to perform well in the wild and not to overfit to training data. On the other hand, obtaining large-scale diverse data in a real-world setting, especially under rare circumstances, is not a simple task. As a consequence, small scale and mostly normal attributes tend to be the common features of the available datasets. Unfortunately, training computer vision models under normal scenarios, such as clear sky, optimal lighting, and standard recording conditions, causes unexpected behaviour or complete failure in much challenging adverse conditions.

In this work, we focus on person tracking under adverse snowy, rainy and foggy weather conditions. The general problem of visual object tracking (VOT) is one of the major tasks in computer vision field that is essential for solving other higher-level tasks such as pedestrian detection, action recognition, or trajectory estimation. Therefore, it is vital for many real-world systems such as self-driving vehicles, automated retail or visual surveillance.

ufuk.celikcan@gmail.com (Ufuk Celikcan), erkut@cs.hacettepe.edu.tr (Erkut Erdem), aerdem@ku.edu.tr (Aykut Erdem)

Failure of such systems under adverse conditions can lead to property damages or human injuries. Thereby, to assess the performance of the state-of-the-art trackers in person tracking in video feeds taken under such circumstances, we collect a novel real dataset, PTAW172Real, that consists of 172 videos featuring weather with heavy snow, rain or fog. Our experiments expose the poor performance of the state-of-the-art trackers when tested on PTAW172Real and this can be linked to the limited number of videos taken under adverse weather conditions in the current VOT datasets that these trackers were trained with. We offer a remedy for the lack of data availability by using our NOVA engine to generate a synthetic dataset, PTAW217Synth, that provides diverse and rich training sequences featuring adverse weather conditions. We show that using synthetic data, we can bridge the aforementioned gap and improve the performance of the learning-based trackers in such conditions. To the best of our knowledge, no work has been done to validate the usability of synthetic data for this purpose.

Our main contributions in this paper can be summarized as follows:

- We present a novel real dataset called PTAW172Real for visual object tracking under adverse weather conditions. The dataset contains 172 videos manually annotated covering snowy, rainy and foggy weather conditions.

- We highlight the poor performance of the state-of-the-art trackers under adverse weather conditions with PTAW172Real.

- Using our NOVA rendering engine, we procedurally generate a new dataset called PTAW217Synth made up of synthetic sequences under adverse weather conditions complete with automatically-generated per-frame annotations including bounding boxes at pixel-level accuracy, occlusion state and other relevant metadata such as time-of-day and camera type. The dataset consists of 217 sequences for person tracking spanning the three adverse weather conditions.

- We show that fine-tuning the pre-trained models on our synthetic dataset PTAW217Synth is able to improve the performance of the deep trackers. Similarly, we also show that training from scratch on only our synthetic training dataset can achieve comparable results to training on large-scale real datasets.

2. Related Work

Despite the relatively short history of deploying synthetically generated data in the field of computer vision, a number of studies have explored the usability of synthetic data for different computer vision tasks. Synthetic data in these studies are utilized for both training and testing purposes. Specifically, for training, it can be used as the sole training data or toward augmenting the actual data in pre-training or fine-tuning of learning models.

Alhaija et al. [7] investigated the use of synthetic data for instance segmentation and object detection. They concluded that training on both synthetic and real data achieves better results as compared to training on a small set of real data. At the same time, they showed that fine-tuning on their augmented data can achieve even better results. Similarly, Cheung et al. [8] proved that synthetic data can be used together with real data to boost accuracy for crowded scene understanding. They showed that using their generated synthetic dataset, LCrowdV, with real datasets can improve the accuracy as compared to using these real datasets alone.

Varol et al. [9] demonstrated the usability of synthetic data for human depth estimation and part segmentation. Their results exhibit that training on synthetic and real images can increase the accuracy for semantic segmentation and reduce the root-mean-squared-error for depth estimation. In the same way, Barbosa et al. [10] extensively studied the advantages of using their generated synthetic dataset, SOMAset, for the task of person ReID. They showed that performing pre-training on their synthetic dataset and then fine-tuning on real datasets achieve better results as compared to training only on real datasets.

Under the scope of VOT, Gaidon et al. [11] provided a detailed analysis on the advantages of using synthetic data for the task of multi-object tracking. Training on their synthetic dataset then fine-tuning on real datasets was shown to achieve the best results as compared to only training on synthetic or real datasets. In the same vein, Zhang et al. [12] used image-to-image translation method to generate synthetic thermal infrared tracking videos using the RGB ones. Their study illustrated that training on their synthetic videos then fine-tuning on real ones or training on both synthetic and real videos achieve better results as compared to training on the available small scale real datasets.

In line with the previous studies, we also investigate the advantages of using synthetic data for training learning-based visual tracking models. How-

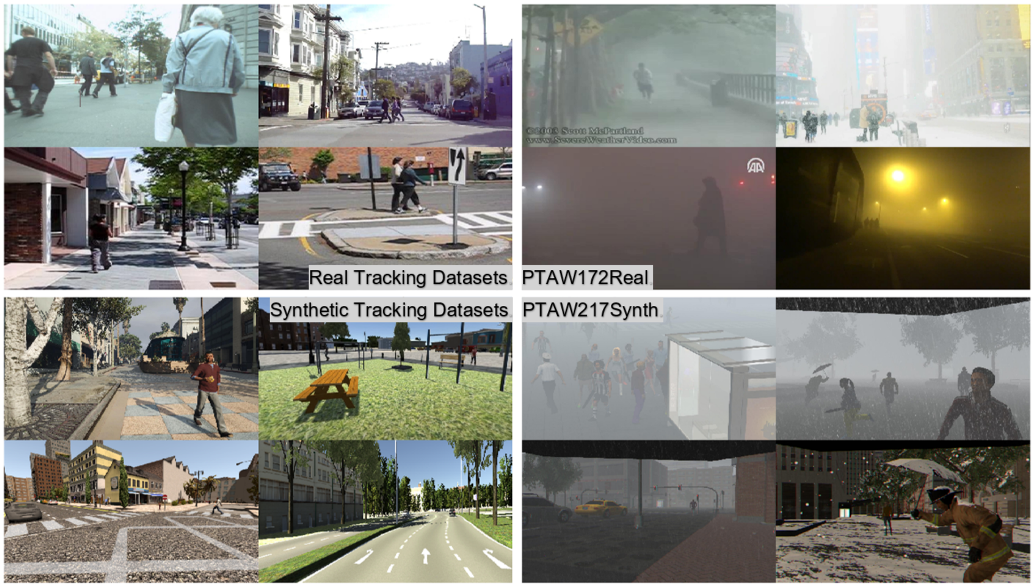

Figure 1: On the left half, sample frames from the currently-available real (top-left quarter) [13, 14, 15, 16] and synthetic (bottom-left quarter) [17, 18, 19, 11] visual object tracking datasets demonstrate the lack of adverse weather conditions. The right half presents sample frames from sequences spanning raining, foggy and snowy weather conditions from PTAW172Real (top-right quarter) and PTAW217Synth (bottom-right quarter) datasets that we introduce in this work.

ever, this work sheds light on the limitations of the existing real and synthetic visual object tracking datasets. As shown in Fig. 1, the adverse weather conditions seem to be underrepresented in most of the available real and synthetic VOT datasets. This causes the state-of-the-art trackers perform poorly under these challenging weather conditions. Bearing this in mind, we present synthetic data as a legitimate solution for the lack of adverse weather conditions in the real datasets. To this end, we utilize our procedural content generation engine NOVA to produce a visual object tracking dataset for training of general purpose visual object trackers. The generated dataset is specifically designed for tracking people under adverse weather conditions in outdoor environments.

3. Extensions to NOVA Framework

To procedurally generate synthetic sequences of pedestrians under adverse weather conditions, we use the NOVA rendering engine [20], which is designed with the goal of allowing researchers with no experience in computer graphics to generate high quality datasets with accurate and dense annotations. NOVA operates in two modes. The first is to generate a single sequence while the other is to output a full dataset. The first mode gives the user full control of the sequence to be generated where it is possible to specify the environment, the weather condition, time of day, camera type, number of cars and number of pedestrians and their density. The dataset mode requires nothing to be specified except the number of sequences to be generated so that NOVA varies the other parameters automatically.

For the particular task of person tracking this work deals with, NOVA generates, for each frame, a bounding box specifying the exact location of the person(s) being tracked in the frame and the occlusion state, that is, whether any other object or person in the scene occludes the person(s) being tracked at that instant. In addition to these, a supplementary metadata is provided with each sequence denoting the environment, weather condition, time of day, camera type, number of people and cars, and people density.

One of the major highlights of NOVA is its capacity to procedurally generate highly diverse and photorealistic sets of synthetic humans. So much so that, each generated human is essentially unique in appearance due to the practically infinite number of recipes (combinations of parameters that are put together randomly on the fly but in cohesion with each other) that NOVA uses in creating them. In this work, we further develop this aspect of NOVA by incorporating premade synthetic humans from Microsoft Rocketbox Avatar Library [21].

Since the main aim of this work is to enhance the performance of visual object trackers under adverse weather conditions, we also extended other capabilities of NOVA toward photorealistic simulation of the generated humans under adverse weather conditions. The environment is built to change dynamically to match the corresponding weather condition and time of the day. For instance, the textures of buildings are changed to have lit windows at nighttime. Furthermore, we implemented the following to facilitate the generation of synthetic sequences with similar visual characteristics to those of the real-world videos captured under adverse weather conditions.

Snowy Weather Condition. First, the variety of clothing used to generate pedestrians in snowy weather is restricted only to outdoor cold-weather clothes. At the same time, the pedestrians are randomly assigned umbrellas, such that, an umbrella is attached to the right or left hand of a pedestrian at random and the animation of the pedestrian is set to match the umbrella mode, *i.e.*, open or close. Snow tracks left by cars and pedestrians are simulated. Furthermore, snow banks and melt snow are created on pavements and roads to give a higher degree of realism. For this, a set of street light poles in the scene are selected at random to determine the positions to place snow banks. Then, from a predefined set of snow banks, one snow bank is instantiated for each position. After that, snow materials are assigned at random to the instantiated snow banks. Following this, the scale and rotation of the snow bank models are randomized to allow for even more diversity. On the other hand, melt snow is simulated by the same snow shader that is used to simulate accumulated snow but with the accumulation parameter set to a random number smaller than the one used for accumulated snow. Making use of the particle system and post-processing effects, falling snow particles and blizzard were randomly introduced to the simulation, as well.

Rainy Weather Condition. Similar to the snowy weather, pedestrians in rainy weather are also generated with outdoor cold-weather clothes; and umbrellas are given to some of them in the same way. In addition, water puddles are simulated to account for water accumulation due to the rain. This is realized by using a puddle shader that is assigned to some of the ground materials (pavements, roads etc.) randomly. For the heavy rain, the rain splash is activated and additional water puddles are instantiated from a set of water puddles. Rain drops are generated using the particle system. Furthermore, rain drops falling on the camera lens are simulated using post-processing effects to match the characteristic of the rainy videos in real life.

Foggy Weather Condition. The clothes of pedestrians produced in foggy weather are not limited to a specific category, but are selected randomly instead. Additionally, fog is simulated using post-processing effects and the Enviro system [22]. Fog density is randomized at run time to give more diversity.

Motion Blur and Chromatic Aberration. These camera effects were simulated additionally to match the image degradations observed in real-life adverse weather videos. Using post-processing, NOVA simulates these two effects procedurally and parametrically. Thus, how severe the effect of these two degradations is randomly configured at run time to provide further

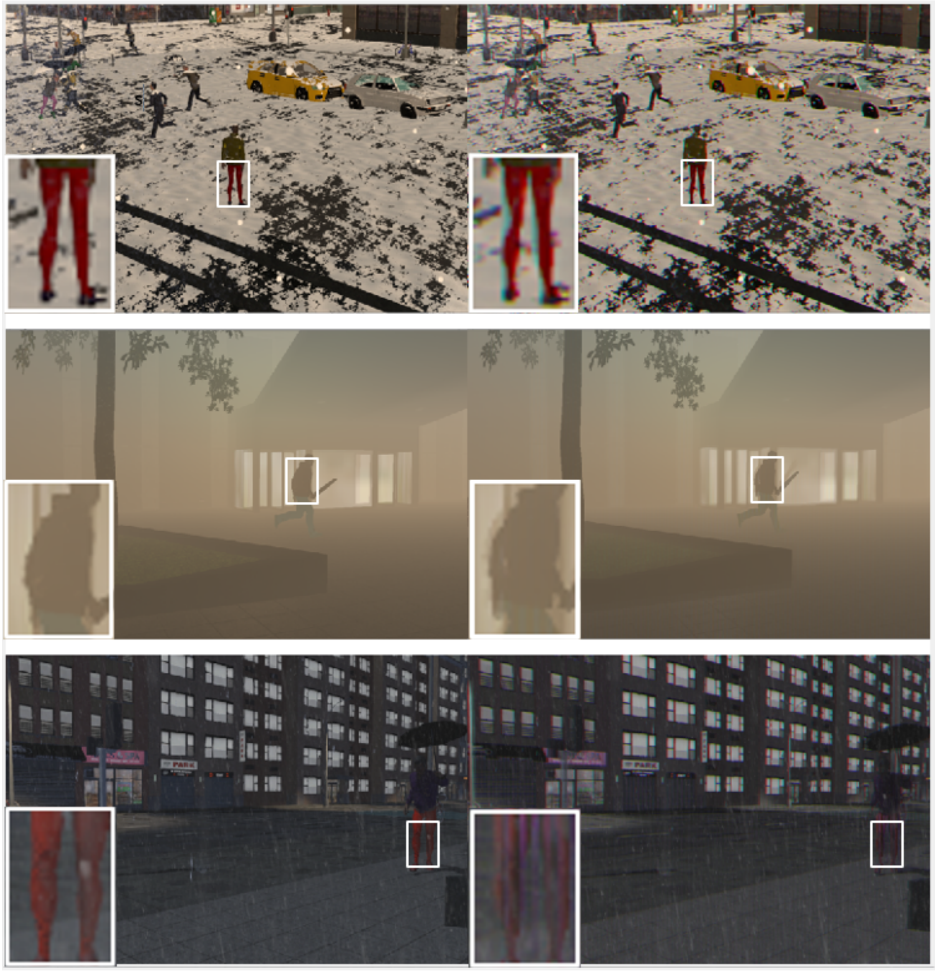

Figure 2: Chromatic aberration, motion blur and both effects are demonstrated in the first, second and third rows, respectively. The first column shows the original frame while the second displays the result of applying the effect(s).

diversity in the generated synthetic sequences. The impact of using these effects over the generated sequences is demonstrated in Fig. 2.

4. PTAW172Real and PTAW217Synth Datasets

4.1. Real-World Data Collection for PTAW172Real

In order to analyze the performance of the recent general purpose visual trackers under adverse weather conditions, we collected real-world videos

Table 1: Dataset statistics of PTAW172Real.

Class	Min Frames	Max Frames	Mean Frames	Total Frames	Videos
Rain	108	1755	498	31888	64
Snow	113	960	394	24010	61
Fog	106	750	328	15394	47
All	106	1755	407	71292	172

from YouTube spanning snowy, rainy and foggy weather. Keywords such as *"adverse"*, *"extreme"*, *"heavy"*, and *"severe"* were used together with the weather names to initiate searches on Youtube. Following this, the query results were manually checked and only the videos satisfying the adverse weather conditions were selected for the annotation. The acquired videos were edited to assure that the person of interest is not occluded and clearly visible in the initial frame. At the same time, the lengths of the videos were modified as needed to keep them around 400 frames per video to provide compatibility with the sequences in the available VOT datasets. Statistics showing the minimum, maximum, average and total number of frames are given in Table 1. The number of videos in the dataset is 172 and the total number of frames is over 71 thousand. The collected videos are at 24 frames per second (FPS) and average time period per sequence is around 17 seconds. Sample frames from the collected PTAW172Real dataset are shown in Fig. 3.

We used the VGG Image Annotator tool [23, 24] for annotating the dataset. We annotated every 5th frame by drawing a bounding box around the person of interest. The tightest box was drawn, excluding the accessories such as handbags, purses, or umbrellas carried by the person. When the person was partially or fully occluded, the estimated location of the person was considered. Additionally, each video was associated with four attributes regarding object occlusion, scale change, background clutter and abrupt camera motion. Fig. 4 gives the hierarchical distribution of the attributes in PTAW172Real dataset.

4.2. Synthetic Data Generation for PTAW217Synth

The PTAW217Synth dataset employed to train the deep learning trackers consists of 217 synthetic sequences that were generated using the NOVA rendering engine. NOVA allows to specify the attributes of the sequences to be generated. In this work, we configured these attributes to match our

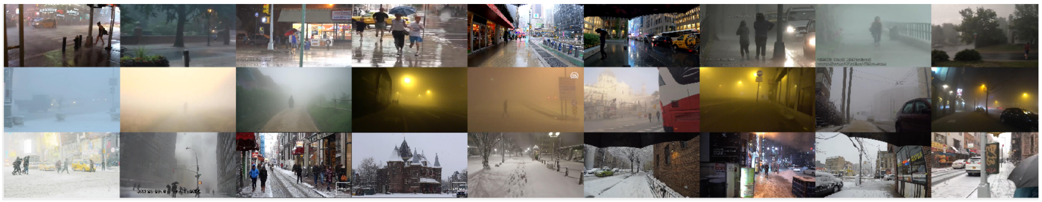

Figure 3: PTAW172Real, our real-world training dataset for person tracking under adverse weather conditions, consists of 172 sequences. Each row shows a specific adverse weather condition, namely, rain, fog, and snow.

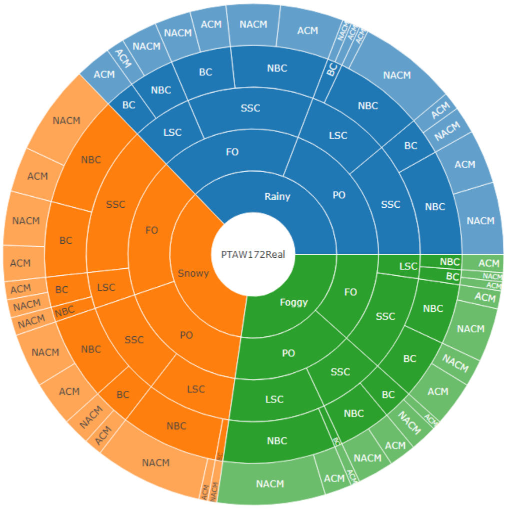

Figure 4: The sunburst chart shows the distribution of different attributes across the PTAW172Real dataset. The innermost circle gives the weather conditions, and the outer circles give occlusion (FO: Full Occlusion, PO: Partial Occlusion), scale change (LSC: Large Scale Change, SSC: Small Scale Change), background clutter (BC: Background Clutter, NBC: No Background Clutter) and abrupt camera motion (ACM: Abrupt Camera Motion, NACM: No Abrupt Camera Motion), respectively.

goal of generating diverse synthetic sequences under adverse weather conditions. Accordingly, the weather conditions were limited to snowy, rainy and foggy weather. The virtual camera type to capture the simulations was set as either the street-level camera or the surveillance camera. The simulation environment was limited to the streets of an urban center, for the reason that such are the most common settings observed in the real-world VOT datasets. In parallel to this, all other attributes, such as time-of-day and crowdedness, were randomized to ensure the diversity of the generated sequences. The

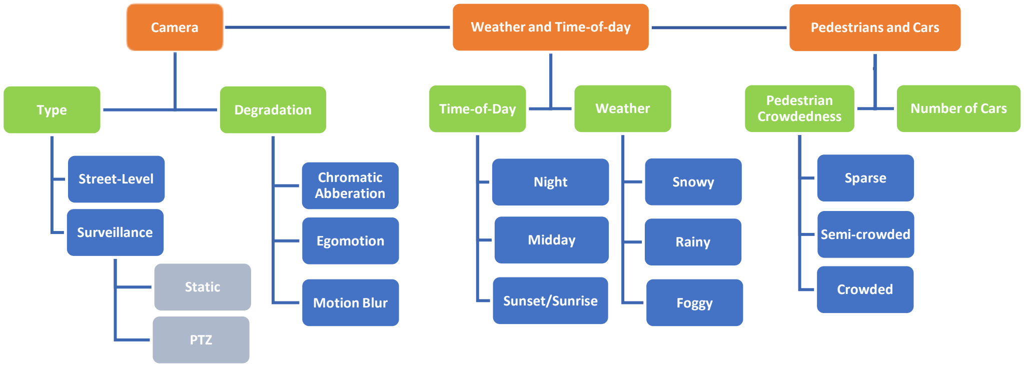

Figure 5: Hierarchical view of the attributes across the PTAW217Synth dataset generated by NOVA.

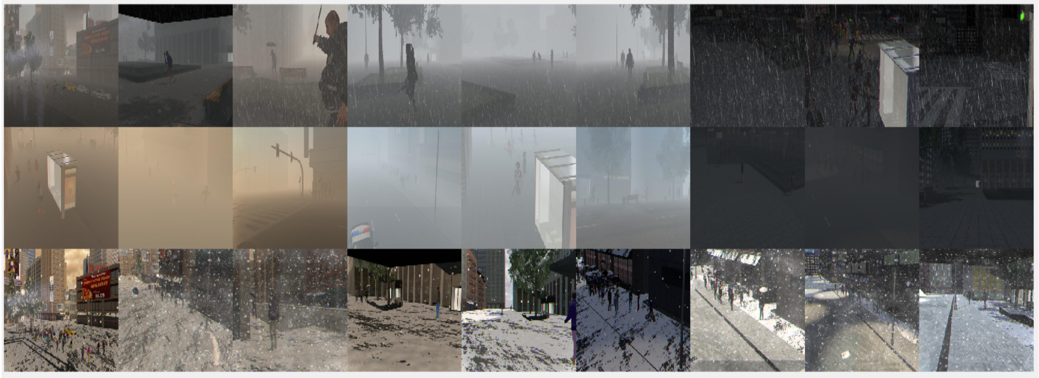

Figure 6: PTAW217Synth, our synthetically generated training dataset for person tracking under adverse weather conditions, consists of 217 sequences, each with a unique set of attributes. Each row shows a specific adverse weather condition, namely, rain, fog, and snow. Sample frames are shown here, illustrating the variations in crowdedness, camera altitude, weather conditions and times of day.

attributes of the generated synthetic sequences are given in Fig. 5. Consequently, the diversity of the generated sequences can be noted in the sample images from these sequences in Fig. 6.

Additional information regarding the minimum, maximum, average and total number of frames are shown in Table 2. The overall average number of frames per sequence is 500 which gives 21-seconds video sequences generated at 24 FPS. The total number of frames of the 217 sequences within the

Table 2: Dataset statistics of PTAW217Synth.

Class	Min Frames	Max Frames	Mean Frames	Total Frames	Videos
Rain	490	510	501	34538	69
Snow	490	510	501	37577	75
Fog	490	510	499	36432	73
All	490	510	500	108547	217

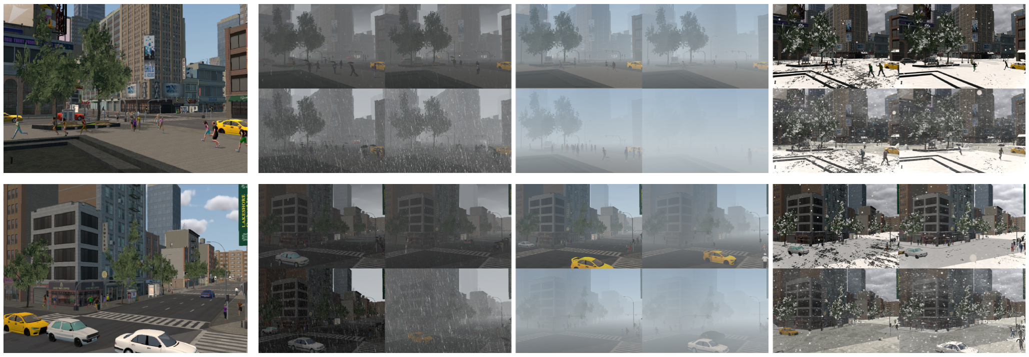

Figure 7: The figure demonstrates the weather variations simulated in PTAW217Synth. The first and second rows present different view points of the same location. Each group of 2x2 images displays a weather condition (from left to right: rainy, foggy, and snowy) in increasing adversity while the large leftmost image in the row shows the same location in clear weather.

dataset is more than 108 thousand. We should note that PTAW217Synth has a balanced distribution of sequences across the rainy, snowy and foggy weather conditions. The sample images captured at a single location from two different view points given in Fig. 7 demonstrate the variety of the simulated weather conditions.

A visual comparison between PTAW172Real and PTAW217Synth datasets is given in Fig. 8. In each row a specific weather condition is presented. Both datasets exhibit similar visual characteristics for the three weather conditions. The figure also demonstrates the level of photorealism of the PTAW217Synth dataset.

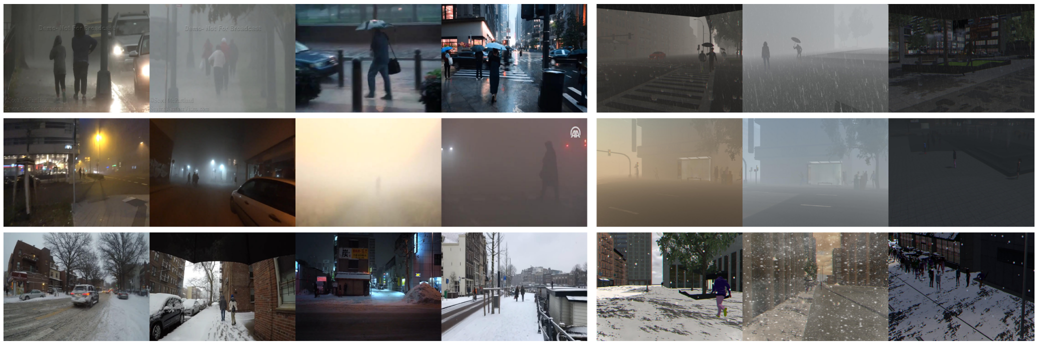

Figure 8: A visual comparison among the synthetic PTAW217Synth (to the right) and real PTAW172Real (to the left) datasets. Each row demonstrates a specific weather condition (from top to bottom: rainy, foggy, and snowy).

5. Experiments

In this section, we study the performance of the state-of-the-art visual trackers in adverse weather conditions in the first set of experiments. The poor performance is highlighted and discussed. In the second set of experiments, we show how the performance of the deep-learning based visual trackers can be enhanced by training on our generated synthetic sequences. First, the evaluation measures are discussed in Section 5.1. Then, the utilized trackers are described in Section 5.2 and the training protocol is explained in Section 5.3. Finally, the results are analysed and explored in Section 5.4.

5.1. Evaluation Measures

The two widely used metrics *precision* and *success* (IoU) are employed for evaluating the performance of the visual trackers analyzed in this work. Precision calculates the distance between the centers of the tracker bounding box and the ground truth bounding box and then checks whether this center error is within the specified limits. We employ the conventional threshold of 20 pixels and consider the tracking as accurate for a frame if the center error is smaller than this value. We then extract the percentage of the accurately predicted bounding boxes for each sequence in our dataset. On the other hand, success measures the intersection over union (IoU) of the tracker and ground truth bounding boxes. We consider a tracking result as successful if the IoU is larger than the common threshold of 0.50, and report the percentage of the successfully predicted bounding boxes averaged over the sequences in our dataset.

5.2. Trackers

In order to properly address the performance of the state-of-the-art general purpose trackers under adverse weather conditions, two different sets of trackers were selected. The sets present the two main approaches in visual object tracking, *i.e.* correlation filter -based and learning-based tracking.

Five state-of-the-art correlation filter -based trackers were chosen for the experiments. These are ECO [25], BACF [26], and context aware (CA) [27] versions of DCF [28], SAMF [29] and STAPLE [30]. DCF, dual correlation filter, utilizes a kernelized correlation filter that has a similar complexity to its linear counterpart, which improves tracker speed (FPS) considerably. On the other hand, SAMF, scale adaptive with multiple features, uses a scale adaptive template size instead of using a fixed one for the correlation filter kernel which is stated to make the tracker more robust. STAPLE, sum of template and pixel-wise learners, fuses template and histogram scores to better handle shape deformation, which facilitates tracking deformable objects more accurately. ECO uses a modified version of DCF to improve memory usage, tracking speed, and robustness. BACF employs a background-aware correlation filter that utilizes specific manually extracted features that account for both background and object of interest change over time. The context aware versions of DCF [28], SAMF [29] and STAPLE [30] that we used improve the original implementations by utilizing the global context information into the standard correlation filter tracking algorithms.

Similarly, for investigating the benefits of training on our generated synthetic sequences, four state-of-the-art learning-based deep trackers were used. These are DiMP [31], ATOM [32], PrDiMP [33], and KYS [34]. DiMP is an offline learning based tracker that can be trained in an end-to-end manner. It applies both background and target information in the process of predicting the object of interest location. The tracker is based on the Siamese tracking architecture. It learns the discriminative loss function during the training phase. ATOM, however, is trained both offline and online. Its tracking algorithm deploys target estimation and classification that are learnt offline and online, respectively. At run-time, the classification component predicts the IoU between the target object and the estimated bounding box. PrDiMP is based on the DiMP architecture. However, unlike DiMP, PrDiMP applies probabilistic regression concept and predicts the probability density of the target given the input frame. This tracker is trained by minimizing KL-divergence in offline manner. KYS tracker uses the visual scene information to better enhance the target localization and tracking. It encodes this information using localized state vectors and propagates it through the sequence to achieve better knowledge of the scene toward realizing better performance during testing. KYS is trained offline to learn how to propagate the scene information.

5.3. Training Protocol

We perform two training scenarios to assess the benefits of the generated synthetic sequences when used for training visual object trackers. For both experiments, the training was done using the whole PTAW217Synth dataset of 217 synthetic sequences. Then, the validation and testing were performed on the whole PTAW172Real dataset. For validation, 33 videos spanning the rainy, foggy and snowy weather conditions were selected at random. The remaining 139 videos were applied for testing.

Training from Scratch. In the first scenario, the models of the four learning-based trackers are trained from scratch using only the generated synthetic sequences. Then, the best model on the validation set is engaged on the test set. Both validation and test sets contain real sequences. The mean and the standard deviation of the tracker performances are reported over 5 iterations to account for the stochastic nature of these trackers.

Fine-Tuning. In the second scenario, the pre-trained versions provided by the authors of the four trackers are fine-tuned on our synthetic sequences. Later, the performances of these models were evaluated on real test sequences as done in the previous scenario.

5.4. Results

The performances in terms of precision and success score are shown in Tables 3 and 4 for the studied trackers on the test partition of PTAW172Real, namely 163 videos. These results show that the trackers from both tracking mainstreams, correlation filter based and learning based, performed poorly under adverse weather conditions. This observation confirms that adverse weather conditions pose certain challenges for the state-of-the-art tracking

Table 3: Precision results of the studied state-of-the-art trackers on the test partition of PTAW172Real, our dataset of real-world outdoor videos taken in adverse weather conditions.

Class	ECO	BACF	STAPLE_CA	SAMF_CA	DCF_CA	ATOM	DiMP	PrDiMP	KYS
Rain	0.59	0.50	0.46	0.38	0.22	0.61+/-0.01	0.60+/-0.01	0.61+/-0.01	0.63+/-0.02
Snow	0.56	0.53	0.49	0.46	0.35	0.60+/-0.01	0.62+/-0.01	0.59+/-0.01	0.58+/-0.01
Fog	0.67	0.65	0.59	0.42	0.37	0.73+/-0.01	0.74+/-0.01	0.74+/-0.01	0.77+/-0.02

Table 4: Success scores of the studied state-of-the-art trackers on the test partition of PTAW172Real, our dataset of real-world outdoor videos taken in adverse weather conditions.

Class	ECO	BACF	STAPLE_CA	SAMF_CA	DCF_CA	ATOM	DiMP	PrDiMP	KYS
Rain	0.64	0.56	0.47	0.45	0.20	0.66+/-0.01	0.63+/-0.01	0.64+/-0.01	0.65+/-0.02
Snow	0.56	0.55	0.49	0.43	0.28	0.59+/-0.01	0.61+/-0.01	0.59+/-0.01	0.57+/-0.01
Fog	0.70	0.69	0.59	0.42	0.27	0.73+/-0.01	0.73+/-0.01	0.73+/-0.01	0.78+/-0.02

algorithms. The correlation filter trackers perform worse than the deep trackers because they are mostly online learning trackers. On the other hand, the deep trackers, which are based on offline learning algorithms, were trained on large scale datasets, which may have contained a number of videos under adverse weather conditions. Thus, they performed slightly better than the ones that are based on correlation filter.

It seems that rain and snow particles, that partially occlude the object of interest, cause a significant change on the visual characteristics handled by the trackers. Thus, it makes it hard for the tracker to differentiate the target object from the background. This effect is particularly clear when the size of the object of interest is relatively small. In parallel to that, fog causes both the background and the object of interest regions to have similar visual appearance. Thus, it makes it hard for the tracker to distinguish the target object from the background. Even so, foggy weather condition seems to be slightly less challenging as compared to the others.

The results of our training experiments are shown in Fig. 9. The IoU scores for the four trained trackers, namely DiMP, ATOM, KYS and PrDiMP, are presented for the two training scenarios. Moreover, these results are compared to the ones of their corresponding baselines. Both the mean and standard deviation were reported over five iterations to account for the stochastic nature of these trackers. Training from scratch on our synthetic adverse weather sequences achieves comparable results to the ones obtained using the baseline for DiMP and PrDiMP. For ATOM and KYS, however, the models trained from scratch surpassed their baselines. On the other hand,

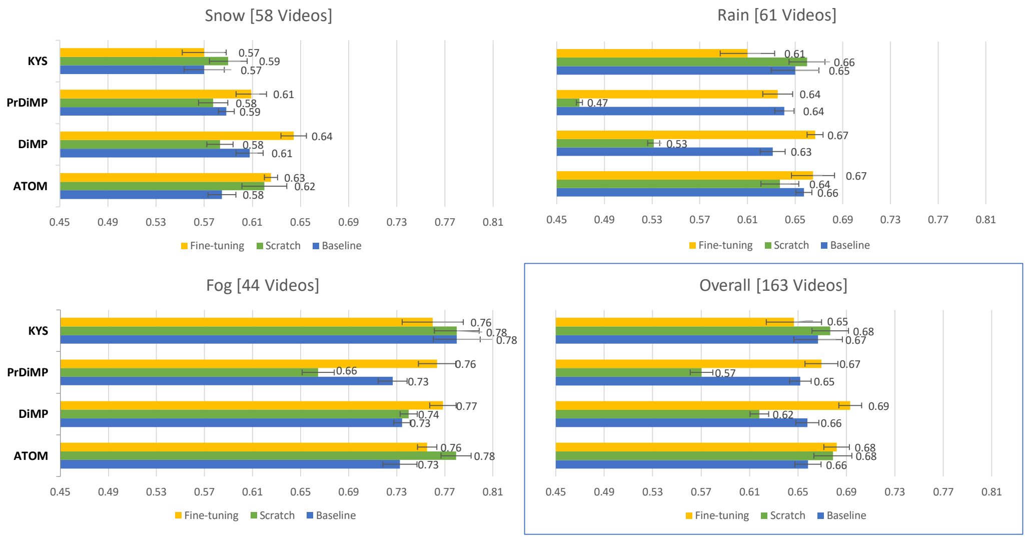

Figure 9: IoU results obtained with the two different training scenarios as compared to those of the baselines. Error bars give the standard deviation of the IoU results.

fine-tuning the pre-trained models on our synthetic sequences improved the performances of ATOM, DiMP and PrDiMP distinctly.

It is worth noting that both the tracking algorithm and the training dataset affect how a specific tracker gains from training on our synthetic sequences. Both determine which training scenario, from scratch or fine-tuning, is more beneficial. For example, DiMP and PrDiMP trackers got the most advantage from fine-tuning. On the other hand, training from scratch was better for KYS tracker, while the performance of ATOM was improved in both scenarios. Another point to be noticed is the conspicuous difference in the level of improvement in trackers performance across different weather conditions. This can be directly linked to the varying distribution of the adverse weather conditions in the different training datasets used for these baselines. So much so that, the lack of adverse weather conditions videos in the training dataset stands out to be the main reason behind the observed performance boost since using even a relatively small number of synthetic sequences spanning these absent features helped the trackers to outperform their baselines, given that the trackers were originally trained on large scale datasets such as LaSOT [35], GOT10k [36], COCO [37], and TrackingNet [38], each far exceeding PTAW217Synth in number of sequences.

It is important to note that the test set contains only real sequences.

Thus, the domain gap problem is not a factor at play under the scope of this analysis. In contrast, diversity of the synthetic sequences in terms of weather conditions, times of day, lighting conditions, camera attributes and synthetic humans altogether enhanced the training process significantly. The high level of photorealism of these synthetic sequences also contributed to diminish the gap between the real and synthetic domains. Thus, performing training from scratch or fine-tuning on our synthetic sequences directly improved the trackers performance.

A qualitative comparison among the tracking results achieved by the baselines and the trained models is presented in Fig. 10. It is seen that utilizing our synthetic data for training improves the performance of the baselines under adverse weather conditions.

Additionally, Fig. 11 displays the success scores for the four deep trackers under full occlusion, scale change, background clutter and sudden camera motion videos. In general, both the baselines and the trained models performed the worst in sequences with background clutter while the ones with sudden camera motion resulted in relatively higher performance. It could be because the background clutter under adverse weather conditions causes the trackers to experience a significant difficulty in locating the object of interest since the background and the object of interest end up having similar visual appearances. On the other hand, the reason that abrupt camera motion does not seem to be effecting trackers as much as the other attributes could be due to the fact that the other three attributes are more closely associated with appearance of the object of interest as compared to the camera motion which translate both the background and the object of interest similarly. A table showing the number of sequences in each weather condition for each of the four attributes is provided in the supplementary material.

6. Conclusion

Our work investigated the lack of adverse weather conditions in the available general purpose visual tracking datasets and highlighted the low performance of the state-of-art trackers in person tracking under these specific circumstances. As a solution, we proposed using our NOVA rendering engine to generate synthetic sequences that span snowy, rainy and foggy weather conditions. We trained four different deep trackers, namely DiMP, ATOM, KYS and PrDiMP, on 217 synthetic sequences generated by NOVA and tested them on the real videos that were collected from YouTube and annotated

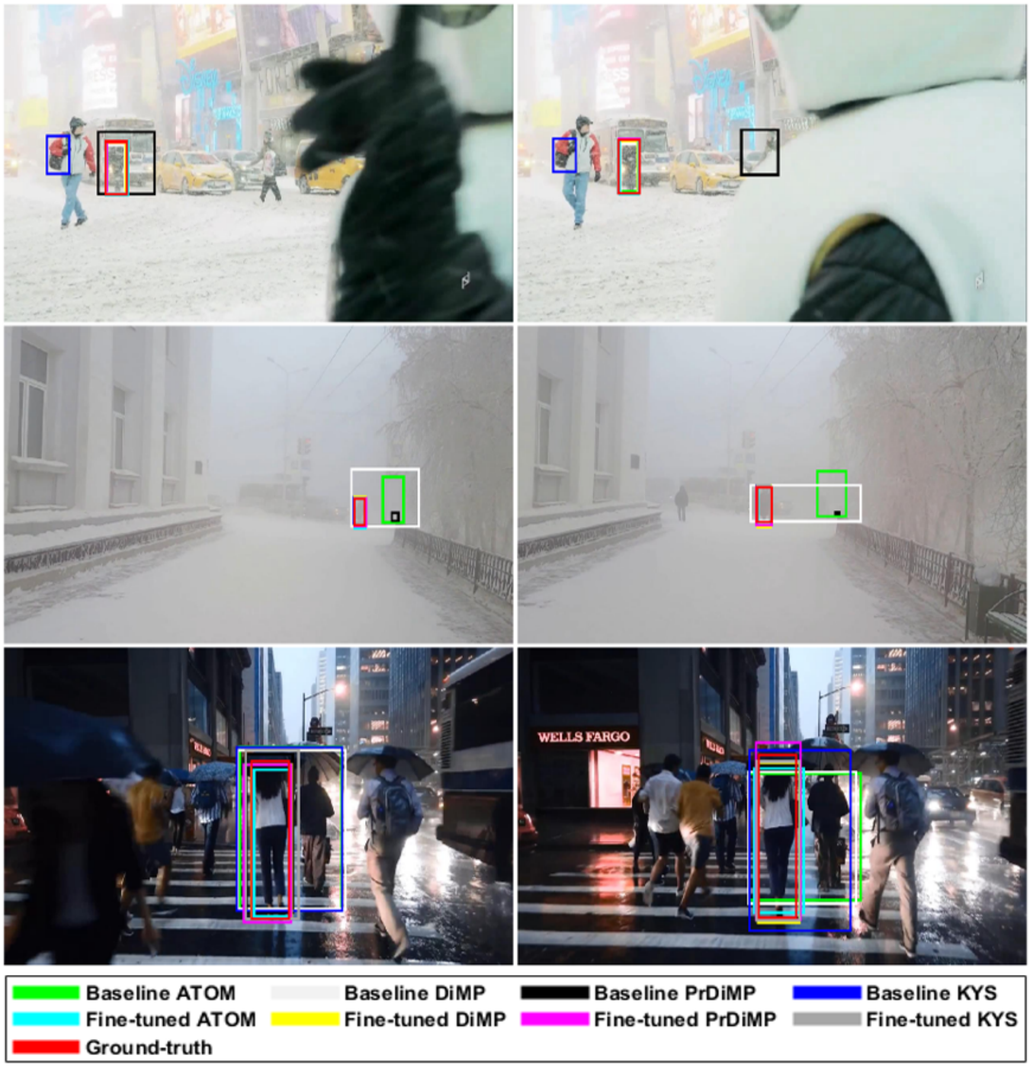

Figure 10: A qualitative comparison of the trained trackers with the baselines on three example sequences. Training on PTAW217Synth improves the trackers performance under adverse weather conditions.

manually for that aim. Our analysis reveals that applying our synthetic sequences for training purposes can bridge the data gap and improve the trackers performance considerably.

A number of limitations have come to light toward the goal of using synthetic sequences for model training as an alternative to the real data. Perhaps

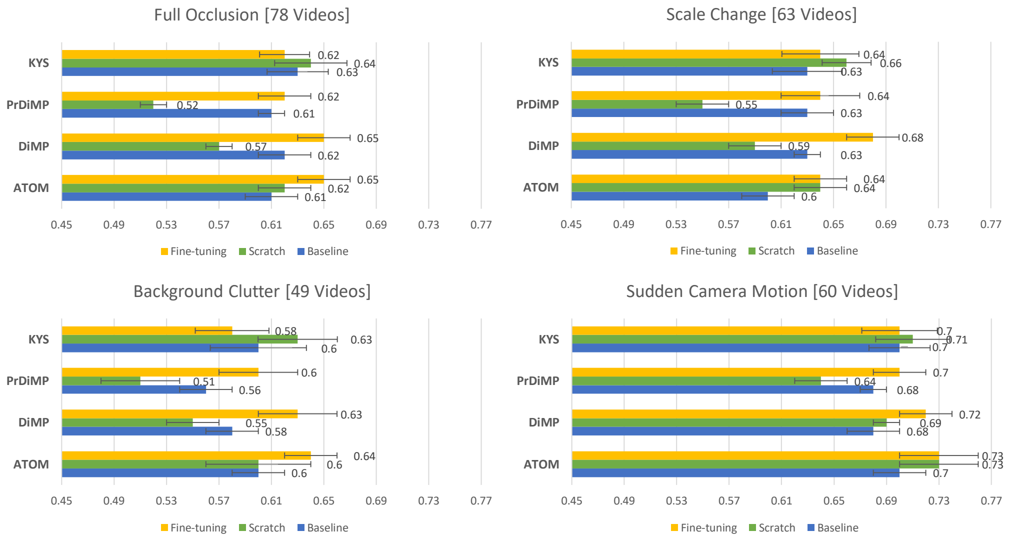

Figure 11: Success scores for ATOM, DiMP, PrDiMP and KYS trackers are shown for four different attributes. Background clutter causes the trackers to perform poorly.

the domain gap problem is the one of central concern in this scope. It arises mainly because the training and testing processes take place in two different domains i.e. synthetic and real domains, respectively. To address this point, we paid great attention to the photorealism of the generated synthetic sequences and most specifically the simulated adverse weather conditions. The second key issue is that synthetic sequences are usually generated at optimal lighting and recording conditions. Thus, the lack of image artifacts such as motion blur, chromatic aberration, noise and others in the synthetic data may cause the models trained on it to fail once such artifacts are encountered in real sequences. To mitigate this problem, we generate our synthetic sequences at different lighting conditions and recording setups. Additionally, we simulate lens artifacts such as motion blur and chromatic aberration. Another note-worthy issue is the fact that repetitive textures, objects, animations, and motions frequently observed in virtual 3D worlds may cause over-fitting. We tackled this issue by diversifying scene elements such as pedestrians, buildings, cars, and other scene objects.

Throughout this work, we demonstrated how our generated synthetic sequences improved trackers performance on adverse weather conditions. However, investigating the effect of adverse weather conditions on other computer vision tasks like optical flow estimation, depth estimation, and person re-identification are sill open questions. The boost in performance upon remedying the lack of sample with adverse weather conditions for the VOT task could be an indication of a similar problem in other computer vision tasks. In light of this study, we believe that using our rendering engine NOVA to generate synthetic training data can bridge the gap of data scarcity in said tasks toward improvement in both accuracy and robustness.

The datasets PTAW172Real and PTAW217Synth that we featured in this work are available for download at the project website `https://graphics.cs.hacettepe.edu.tr/NOVA-Adverse` along with a supporting video illustrating the motivation behind this work, a sample of sequences from PTAW217Synth and also a sample of the PTAW172Real sequences superimposed with tracking results.

Acknowledgements

This work was supported in part by TUBITAK-1001 Program (Grant No. 217E029), GEBIP 2018 fellowship of Turkish Academy of Sciences awarded to E. Erdem, and BAGEP 2021 Award of the Science Academy awarded to A. Erdem.

References

[1] Z. Cai, N. Vasconcelos, Cascade r-cnn: Delving into high quality object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6154–6162.

[2] Q. Chu, W. Ouyang, H. Li, X. Wang, B. Liu, N. Yu, Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4836–4845.

[3] C. Hazirbas, L. Ma, C. Domokos, D. Cremers, Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture, in: Asian conference on computer vision, Springer, 2016, pp. 213–228.

[4] K. Tateno, F. Tombari, I. Laina, N. Navab, Cnn-slam: Real-time dense monocular slam with learned depth prediction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6243–6252.

[5] J. Walker, A. Gupta, M. Hebert, Dense optical flow prediction from a static image, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2443–2451.

[6] J. K. Kang, T. M. Hoang, K. R. Park, Person re-identification between visible and thermal camera images based on deep residual cnn using single input, IEEE Access 7 (2019) 57972–57984.

[7] H. A. Alhaija, S. K. Mustikovela, L. Mescheder, A. Geiger, C. Rother, Augmented reality meets computer vision: Efficient data generation for urban driving scenes, International Journal of Computer Vision 126 (9) (2018) 961–972.

[8] E. Cheung, T. K. Wong, A. Bera, X. Wang, D. Manocha, Lcrowdv: Generating labeled videos for simulation-based crowd behavior learning, in: European Conference on Computer Vision, Springer, 2016, pp. 709–727.

[9] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, C. Schmid, Learning from synthetic humans, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 109–117.

[10] I. B. Barbosa, M. Cristani, B. Caputo, A. Rognhaugen, T. Theoharis, Looking beyond appearances: Synthetic training data for deep cnns in re-identification, Computer Vision and Image Understanding 167 (2018) 50–62.

[11] A. Gaidon, Q. Wang, Y. Cabon, E. Vig, Virtual worlds as proxy for multi-object tracking analysis, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4340–4349.

[12] L. Zhang, A. Gonzalez-Garcia, J. van de Weijer, M. Danelljan, F. S. Khan, Synthetic data generation for end-to-end thermal infrared tracking, IEEE Transactions on Image Processing 28 (4) (2018) 1837–1850.

[13] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, K. Schindler, Mot16: A benchmark for multi-object tracking, arXiv preprint arXiv:1603.00831 (2016).

[14] A. Li, M. Lin, Y. Wu, M. Yang, S. Yan, NUS-PRO: A New Visual Tracking Challenge, IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (2) (2016) 335–349.

[15] Y. Wu, J. Lim, M. Yang, Object tracking benchmark, IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (9) (2015) 1834–1848. doi:10.1109/TPAMI.2014.2388226.

[16] P. Liang, E. Blasch, H. Ling, Encoding color information for visual tracking: Algorithms and benchmark, IEEE Transactions on Image Processing 24 (12) (2015) 5630–5644.

[17] S. R. Richter, Z. Hayder, V. Koltun, Playing for benchmarks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2213–2222.

[18] C. R. De Souza, A. Gaidon, Y. Cabon, A. M. L. Peña, Procedural generation of videos to train deep action recognition networks., in: CVPR, 2017, pp. 2594–2604.

[19] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, A. M. Lopez, 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, 2016, pp. 3234–3243.

[20] A. Kerim, C. Aslan, U. Celikcan, E. Erdem, A. Erdem, Nova: Rendering virtual worlds with humans for computer vision tasks, Computer Graphics Forum, in press (2021).

[21] Microsoft, Microsoft rocketbox avatar library git repo, https://github.com/microsoft/Microsoft-Rocketbox, online; accessed: 2020-05-17.

[22] P. Worlds, Enviro webpage, online; accessed: 2019-02-20.
URL http://www.procedural-worlds.com/gaia/gaia-extensions/enviro/

[23] A. Dutta, A. Zisserman, The VIA annotation software for images, audio and video, in: Proceedings of the 27th ACM International Conference on Multimedia, MM '19, ACM, New York, NY, USA, 2019. doi:10.

1145/3343031.3350535.
URL https://doi.org/10.1145/3343031.3350535

[24] A. Dutta, A. Gupta, A. Zissermann, VGG image annotator (VIA), http://www.robots.ox.ac.uk/ vgg/software/via/, version: 2.0.10, Accessed: 2020-07-19 (2016).

[25] M. Danelljan, G. Bhat, F. Shahbaz Khan, M. Felsberg, Eco: Efficient convolution operators for tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6638–6646.

[26] H. Kiani Galoogahi, A. Fagg, S. Lucey, Learning background-aware correlation filters for visual tracking, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1135–1143.

[27] M. Mueller, N. Smith, B. Ghanem, Context-aware correlation filter tracking, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[28] J. F. Henriques, R. Caseiro, P. Martins, J. Batista, High-speed tracking with kernelized correlation filters, IEEE transactions on pattern analysis and machine intelligence 37 (3) (2015) 583–596.

[29] Y. Li, J. Zhu, A scale adaptive kernel correlation filter tracker with feature integration, in: European conference on computer vision, Springer, 2014, pp. 254–265.

[30] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, P. H. Torr, Staple: Complementary learners for real-time tracking, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1401–1409.

[31] G. Bhat, M. Danelljan, L. V. Gool, R. Timofte, Learning discriminative model prediction for tracking, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6182–6191.

[32] M. Danelljan, G. Bhat, F. S. Khan, M. Felsberg, Atom: Accurate tracking by overlap maximization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4660–4669.

[33] M. Danelljan, L. V. Gool, R. Timofte, Probabilistic regression for visual tracking, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7183–7192.

[34] G. Bhat, M. Danelljan, L. Van Gool, R. Timofte, Know your surroundings: Exploiting scene information for object tracking, arXiv preprint arXiv:2003.11014 (2020).

[35] H. Fan, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, C. Liao, H. Ling, Lasot: A high-quality benchmark for large-scale single object tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5374–5383.

[36] L. Huang, X. Zhao, K. Huang, Got-10k: A large high-diversity benchmark for generic object tracking in the wild, IEEE Transactions on Pattern Analysis and Machine Intelligence (2019).

[37] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755.

[38] M. Muller, A. Bibi, S. Giancola, S. Alsubaihi, B. Ghanem, Trackingnet: A large-scale dataset and benchmark for object tracking in the wild, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 300–317.