Temporal and Non-Temporal Contextual Saliency Analysis for Generalized Wide-Area Search within Unmanned Aerial Vehicle (UAV) Video

Simon G. E. Gökstorp[1] and Toby P. Breckon[1,2]

***Abstract*— Unmanned Aerial Vehicles (UAV) can be used to great effect for wide-area searches such as search and rescue operations. UAV enable search and rescue teams to cover large areas more efficiently and in less time. However, using UAV for this purpose involves the creation of large amounts of data, typically in video format, which must be analysed before any potential findings can be uncovered and actions taken. This is a slow and expensive process which can result in significant delays to the response time after a target is seen by the UAV. To solve this problem we propose a deep model using a visual saliency approach to automatically analyse and detect anomalies in UAV video. Our Temporal Contextual Saliency (TeCS) model is based on the state-of-the-art in visual saliency detection using deep convolutional neural networks (CNN) and considers local and scene context, with novel additions in utilizing temporal information through a convolutional LSTM layer and modifications to the base model. We additionally evaluate the impact of temporal vs non-temporal reasoning for this task. Our model achieves improved results on a benchmark dataset with the addition of temporal reasoning showing significantly improved results compared to the state-of-the-art in saliency detection.**

I. Introduction

Modern advances in technology have enabled the use of Unmanned Aerial Vehicles (UAV) for the purposes of surveillance and search and rescue operations, reducing the costs and improving the capabilities of such operations. UAV can cover large distances and areas quickly and efficiently; however, processing and analysing the video recorded by UAV is still a costly and time-consuming task. The time to response is often critical to the outcome of search and rescue operations, meaning an automated solution which reduces the cost and increases the speed of this analysis would be beneficial for this task.

Visual saliency is a measure of the conspicuity of objects in an image, meaning how much they stand out from the image or how unique they are [1]. Through the application of visual saliency detection, computer vision systems are capable of identifying and extracting the most distinctive parts of an image. Contextual saliency is an extension of visual saliency which considers the context of an image in determining the salient objects, or anomalies, in it. There are various types of context which can be considered such as the local context (pixel neighbourhood) of a given pixel, or the type of scene portrayed by the image. In processing videos as a whole instead of images individually, video saliency detection approaches seek to apply temporal reasoning to improve the accuracy and consistency of saliency predictions, typically by propagating information from previous frames to be considered when processing future frames. By applying these concepts to the task of anomaly detection in UAV video the goal is to produce a general solution which is capable of detecting any object of interest in the video.

Images and video captured by UAV commonly feature a set of distinct properties when compared to the images considered in most saliency detection approaches. These include potentially being grainy, low-quality and noisy (from the motion of the UAV, encoding/transmission of the image, etc.), the possibility of being captured at varying altitudes (and thus scales) and speeds (and thus level of motion noise) and more. Additionally, the salient objects in typical images considered for saliency detection are often large in the image and placed at or near the centre. However, the salient objects in UAV images are typically very small and scattered across the image. These properties place limitations on the application of general saliency detection methods to UAV images, but may also be exploitable by a model specifically designed for this type of imagery.

Previous approaches to the problem of anomaly detection in UAV video have commonly relied on classical computer vision techniques to achieve saliency detection, for example colour space manipulation in [2] and image region segmentation in [3]. Those approaches achieving the best results are generally very slow, taking several minutes to process each frame, and they often do not scale well with larger image sizes [4], making them unsuitable for processing UAV data where target objects may be very small in the image. More recent approaches have achieved great results while limiting the scope of the solution to detecting a small set of object classes, or only considering a specific target environment. Previous approaches have also disregarded the temporality of video by processing frames independently, making them more versatile but less suited for processing video.

In order to solve this problem we evaluate the benefit of temporal information processing for anomaly detection in UAV video, and propose a novel Temporal Contextual Saliency (TeCS) model based on the Deep Spatial Contextual Long-term Recurrent Convolutional Network (DSCLRCN) model of [5], part of the state-of-the-art in saliency prediction using a deep convolutional neural network (CNN) approach. Our proposed model considers local and scene context in each frame, and is novel in leveraging the temporal informa-

[1]Department of Computer Science, Durham University, Durham, UK
[2]Department of Engineering, Durham University, Durham, UK

tion in UAV video.

While the state-of-the-art in saliency prediction has recently been dominated by deep learning approaches, no such approach has previously been applied to the task of salient object detection in UAV video. Within this work we detail novel additions to the baseline DSCLRCN architecture proposed by [5], and additionally explore the use of temporal vs. non-temporal reasoning within a further extended architectural approach. Specifically, we evaluate the impact of using a convolutional Long Short-Term Memory (convLSTM) layer in place of a standard convolution operator on overall saliency detection across a number of exemplary UAV missions (video episodes) and show improved benchmark performance on the UAV123 dataset [6].

II. Related Work

One of the first and most seminal works on visual saliency detection is [1], which has served as the basis and inspiration of many more recent methods such as [7] and [8]. These works use a bottom-up approach based on low-level features such as intensity, colour and orientation, inspired by neuroscience principles. Due to the focus on low-level information, these approaches commonly suffered shortcomings such as reliance on priors, difficulties in detecting objects that touch the edges of the image and in detecting smaller and more subtle objects. Additionally, the approach of [8] suffered from over-detection in UAV/aerial-style images.

Other approaches considered high-level information in the image in the form of the context of the image. This is information about the general contents of the image as a whole, for example the terrain, environment or conditions displayed in the image, or the presence of additional objects in other areas of the image. One of the earliest usages of context for automated saliency detection is [9], which utilized an "auto-context classifier" to learn the context of a salient object through a prior step of iterative learning. More recently, [5] sought to use contextual information together with a neural network based approach for saliency detection, proposing the DSCLRCN model. This model evaluates saliency per pixel in the image while considering the local, global and scene context, achieving better results than all previous models on eye-fixation datasets.

Another neural network based approach was presented by [10], which achieved significantly faster processing speeds by using a fully convolutional network. However, this approach was not designed for nor tested on UAV footage, and resizes images to 352×352 for evaluation, potentially losing out on small-scale information and context which could be very important for UAV images. It also did not consider the scene context of the image unlike [5], instead processing only local and global context within the image.

Early methods specifically designed for salient object detection in UAV imagery such as [2] and [3] were commonly based on the bottom-up approach of [1]. These methods achieve good results by targeting specific scenarios, such as *"rural, uncluttered and relatively uniform environments"* [2] and detecting people and vehicles on roads [3].

Very recently a survey of UAV saliency detection carried out by [4] was built upon by [11]. Based on their findings, [11] present an approach that uses the wavelet transform-based model in [12] to produce a saliency map which is used to select the 300 most salient patches in the image. Next, a CNN trained to detect people is applied to each patch. Their model achieves state-of-the-art results, achieving a higher precision but lower recall score than a Faster R-CNN model [13] trained on the same dataset. However, the model is only designed for the scope of detecting people in land-based situations and is therefore not directly generalizable to the more general task of anomaly detection, and it does not utilize temporal information.

Considering temporal information could massively benefit any saliency approach that is designed for video. A model for video saliency prediction for non-UAV videos is presented by [14], which utilizes a deep CNN and spatial-temporal object candidates to improve the temporal consistency of the saliency prediction. Another approach was taken by [15], who used the convLSTM architecture created by [16] to process spatial-temporal information in video bidirectionally. No previous approaches were found which utilize temporal information to process UAV video.

Previous methods for saliency detection in UAV images and video are generally limited in scope, not considering contextual or temporal information available, or making assumptions about the type of salient object or environment expected. While there has been a large amount of research into the topics of contextual saliency and video saliency, these ideas have not been extensively applied to UAV video. In the field of visual saliency detection deep learning models are dominating the state-of-the-art, both in terms of accuracy and execution speeds. A recent evaluation of the performance of existing visual saliency models on UAV video by [17] drew the same conclusion, while stressing the importance of developing UAV-centric models tailored for this task.

Our proposed TeCS model is novel in applying these ideas to the topic of anomaly detection in UAV video. It does so by building on the DSCLRCN model of [5]. By adapting this model by replacing the last convolutional layer with a convolutional LSTM layer and changing the activation function of the last convolutional layer as well as the loss function we produce our novel TeCS model, which achieves significantly improved salient object detection performance in UAV video compared to the base DSCLRCN model. A comparison of a temporal and non-temporal version of this model demonstrates the significant improvement yielded by temporal processing.

III. Solution

Our proposed solution is a deep CNN model based on the state-of-the-art in contextual saliency detection. The model is adapted to the task of anomaly detection in UAV video by changing the activation function as well as the loss function used to train the model. It additionally utilizes temporal information carried in video by propagating data through

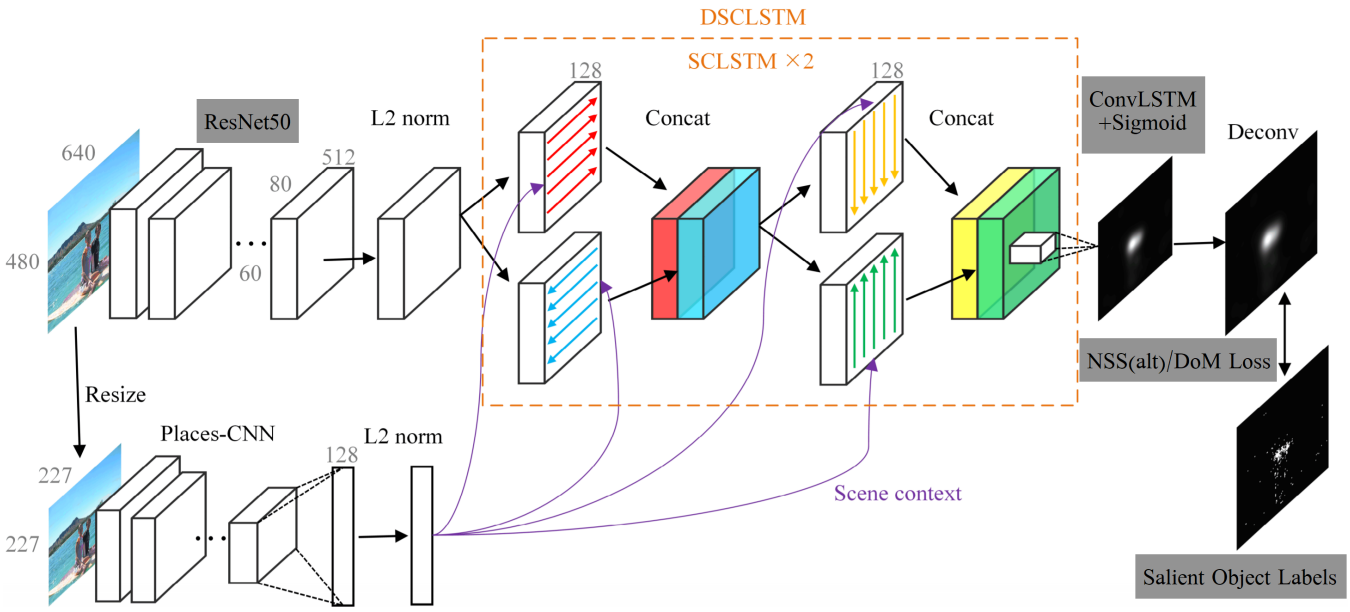

Fig. 1. An overview of our proposed architecture - original DSCLRCN (white background, taken from [5]) and modifications (light grey background, our variant TeCS).

time to improve the analysis of subsequent frames via a convolutional LSTM layer.

Based on the results of the literature survey, we chose to construct the solution based on the state-of-the-art deep learning model for contextual saliency proposed by [5]. This choice was made because the survey of related works revealed that deep learning models generally outperform classical computer vision approaches, both in terms of accuracy and execution speeds. The structure of our proposed TeCS model is shown in Fig. 1. For more details of the original architecture see [5].

A. Modifications for UAV Data

In order to adapt the DSCLRCN model for use with UAV images we make several modifications to the model architecture and training procedure. Firstly, we change the activation function applied to the output of the final convolution layer, originally the *Softmax()* function, to the *Sigmoid()* function. Although the lateral competition introduced by the *Softmax()* function is desirable as it helps produce cleaner saliency predictions, it has the side-effect that the magnitude of the output is always the same. A model using the *Softmax()* activation function as the last activation function is therefore unable to produce an output that contains no predicted saliency for an input image. The model is also incapable of predicting the overall saliency level of an image (i.e. whether the image contains many or very few salient objects, the magnitude of the saliency prediction remains the same).

This is not an issue for the case of typical visual saliency prediction, as the model should predict the most salient item in every image. Such cases therefore have no negative examples (images with no salient objects in them). This is however an issue for applying saliency prediction for salient object detection, as the model should be able to predict a lack of any salient objects in an image. Using the *Sigmoid()* activation function removes this issue. As this function has a range of $(0, 1)$, it is well suited for tasks that evaluate probabilities. By applying this activation function to the output of the last convolutional layer, each pixel in the output is assigned a value in this range, corresponding to the saliency of that pixel. As the *Sigmoid()* function is applied to each pixel individually, no constraints are placed on the image as a whole, or on the relationship between pixels. The model is thus able to output a low value at every pixel in the image if it does not detect any salient objects.

We also adjust the testing procedure used when validating and testing the model. The authors of [5] found that applying a Gaussian blur to the saliency prediction produced by the DSCLRCN model improved its performance by smoothing out the saliency response. Such blurring may improve the saliency prediction for large objects by removing large peaks and small gaps in the prediction but it also removes detail at smaller scales. In UAV video target salient objects can be present in varying scales due to factors such as the altitude of the UAV. We therefore omit this stage of processing in order to preserve small-scale detail in the predictions.

In addition to the changes made to the architecture and post-processing of the model we also change the loss function used to train the model. To train the DSCLRCN model [5] used the negative Normalized Scanpath Saliency (NSS) [18] to compute the loss of a prediction with respect to the ground truth from human eye fixation data. However, the NSS loss function assumes the presence of target pixels in the ground truth. If there are no targets in the ground truth fixation data, as could be the case in the data considered for UAV anomaly detection, then the NSS is not defined. Therefore, we are unable to use this loss function for training our model while including images that contain no salient object in the

dataset. Another loss function commonly used in saliency prediction is Pearson's Correlation Coefficient (CC), which was recommended for use for saliency prediction evaluation by [19]. This function suffers from the same problem as the NSS score, being undefined for images where the ground truth has no salient objects, and thus is also unsuitable.

In order to solve this problem, we investigated several other loss functions for training our model. First, based on the recommendation of [15] we used a compound loss function of the Cross Entropy (CE) and the Mean Absolute Error (MAE), CE_MAE, of the predicted saliency compared to the ground truth. By combining these two loss functions in this way, [15] found that their model for video salient object detection achieved better results as the compound loss function better captured different factors contributing to the overall quality of the results.

A second loss function we investigated was a modified version of Normalized Scanpath Saliency. We noted that the NSS loss function had been used to great success in recent works, and is recommended for evaluating saliency predictions by many surveys of common metrics such as the work of [19] and [20], which found that out of nine scores surveyed NSS performed the most consistently with human evaluations. For these reasons, we wished to apply the NSS loss function to our task of anomaly detection in UAV video, while still being able to include negative images in the dataset. Our chosen approach for this was to use the NSS loss function when possible, and apply a different loss function when the NSS is not defined. Given a prediction x and ground truth y, the resultant NSS_{alt} loss function is computed as:

$$NSS_{alt}(x,y) = \begin{cases} -\frac{\sum(\bar{x} \circ y)}{\sum y} & \sum y > 0 \\ \sigma(x) & \sum y = 0 \end{cases} \quad (1)$$

$$\bar{x} = \frac{x - \mu(x)}{\sigma(x)} \quad (2)$$

where $\circ$ denotes element-wise product, $\bar{x}$ is the saliency map of x normalised to have a mean of 0 and standard deviation of 1, μ denotes the mean of x, and σ denotes the standard deviation of x. The rationale behind the design of this function is that if there is no target salient object in the ground truth y, then the model should output a predicted saliency map that is monotonous and invariable across the image, as there are no spatial locations in the image that are more salient than the others. Although this loss function is likely imperfect, and is not well balanced between the two cases as the ranges of them are significantly different, this simple alteration allows us to apply the NSS loss function to our UAV data.

We also considered another loss function which we created, inspired by the Normalized Scanpath Saliency function. We took the idea of NSS to measure the mean predicted saliency value at target salient points, but rather than normalising the saliency prediction to a mean of 0 and a standard deviation of 1, we introduce a second term in the form of the mean predicted saliency value at non-target points. This loss function, which we name Difference of Means (DoM), is computed as:

$$DoM(x,y) = \mu(x_i, y_i = 0) - \mu(x_i, y_i > 0) \quad (3)$$

where $\mu(x_i, y_i = 0)$ denotes the mean value of the set of pixels in x where the corresponding location in y has a value of 0. If no pixel in y has a value greater than 0, $\mu(x_i, y_i > 0)$ is taken to be 0. The investigation of this loss function was inspired by the observation that the dataset used in training our model for anomaly detection in UAV video contained a large number of frames with a single small target. This meant that when trained with some loss functions such as CE_MAE recommended by [15] the model was able to achieve a very low error by outputting low saliency predictions throughout the image. This issue led us to want a loss function where the task of predicting high saliency at the salient object locations and the task of predicting low saliency at non-salient locations were balanced, rather than each pixel being treated as equal. Additionally, this loss function has an advantage in that it is applied equally to all images and ground truths, unlike the NSS_{alt} loss function which uses a piecewise function to handle ground truths with no salient objects.

In order to speed up the learning process we use the Adam optimiser [21] with a learning rate of 0.01, a β_1 of 0.9 and a β_2 of 0.999. When training our non-temporal model we instead use SGD with a momentum of 0.9 and weight decay of 0.0005, as per in [5]. We also use a learning rate scheduler to reduce the learning rate by a factor of 2.5 every epoch. Since pre-trained weights are used for the local feature extractor and the scene context extractor models we reduce the learning rates for these layers by a factor of 0.1 compared to the rest of the model, allowing the weights to be fine-tuned for our task and reducing the risk of decay in performance of these parts of the model. Implementing the above modifications produces the non-temporal version of our proposed TeCS model, NTeCS.

B. Temporal Implementation

As the DSCLRCN model is designed for the task of visual saliency prediction in images, it is not adapted to processing videos. We therefore further augment the model to leverage the temporal consistency of the saliency in consecutive frames of a video, producing our proposed TeCS model. We do this by replacing the final convolution that reduces the channel dimension to 1 for saliency prediction with a convolutional LSTM (convLSTM) layer [16]. By using a convLSTM layer, the saliency prediction at each spatial location is computed as a function of the feature vector computed by the previous layer in the model at that location and neighbouring locations, as well as feature vectors from previous frames at that location and neighbouring locations.

We apply a convLSTM layer with 3×3 kernel size and 256 input channels and a single output channel. As the output is produced using the *tanh()* activation function, which has a range of $(-1, 1)$, the output values cannot be directly output

as saliency prediction values. Since *tanh()* is a rescaled *Sigmoid()* function, we map the output of the convLSTM layer h_t to the range $(0, 1)$ as $p_t = \frac{h_t+1}{2}$. After deconvolution we threshold the output to produce the saliency prediction p.

C. Dataset

There is currently no publicly available dataset designed for the task of salient object detection in UAV video. Due to this, we use the UAV123 dataset [6] to train, validate and test our proposed model. Although this dataset is designed and labelled for object tracking, not salient object detection, a significant number of the sequences in it feature a single salient object and thus the ground truth data function well as salient object labels. We also considered a subset of this dataset labelled for human visual attention named EyeTrackUAV, created by [22]. However, the original labels serve better as salient object labels which we need for our task, and therefore we do not use this dataset.

In order to improve the quality of the dataset for use for our task we remove all 'building', 'UAV' and 'bird' sequences due to their design and the extreme levels of noise present. We also removed all sequences produced by simulation, leaving a total of 70 sequences. We split the sequences into training, validation and testing sets with 35, 17 and 18 sequences respectively. We spread sequences with the same class of target object as evenly as possible between the sets. Due to the large total number of frames in the dataset we only use the first 300 frames of each sequence, resulting in $\sim$10000 total frames in the training set and $\sim$5000 frames each in the validation and testing sets. This was done to reduce the training time of the model without further reducing the number of different sequences considered.

TABLE I

PERFORMANCE OF MODELS ON OUR UAV123 [6] TEST SET.

Architecture	$\uparrow$NSS$_{alt}$(+)	$\downarrow$NSS$_{alt}$(-)	$\downarrow$CE_MAE	$\downarrow$DoM
DSCLRCN	3.552	0.091	0.286	-0.398
NTeCS	3.315	0.163	0.220	$\mathbf{-0.571}$
TeCS	**8.851**	**0.023**	**0.144**	-0.251

IV. EVALUATION

We compare three distinctive model architectures: DSCLRCN, the baseline, NTeCS, our proposed solution without the temporal implementation, and TeCS, our full proposed model. We report the results of each model using several loss functions as performance metrics: our NSS$_{alt}$ score, which was used to train the TeCS model, split into positive and negative images, Cross Entropy and Mean Absolute Error (CE_MAE) based on the recommendation of [15], and our DoM score, which was used to train the NTeCS model. All models were tested using a GeForce RTX 2080 Ti GPU, and run at a processing speed of 2.2 FPS without any parallel processing. The NTeCS model was trained using the SGD optimiser while Adam was used for the TeCS model. Each model was trained for 10 epochs, with validation experiments after each epoch.

An overview of the performance of the different models on the test set we created of UAV123 sequences is shown in Table I. In this table the NSS$_{alt}$ metric is reported separately for images containing some salient pixels and images containing none, indicated by (+) and (-), respectively. Additionally, for each metric the arrow indicates whether a higher or lower score is better, and the best score for each is shown in bold. These results clearly show that our proposed

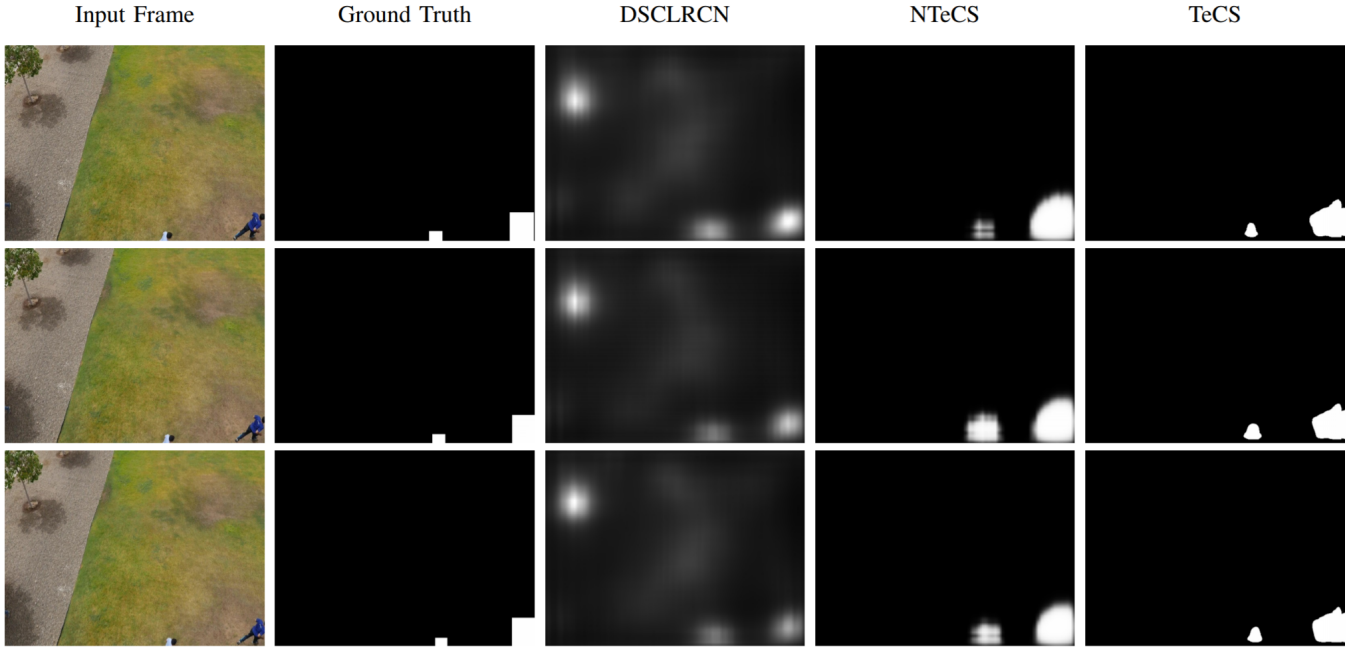

Fig. 2. Performance of models on the 'person9' sequence from UAV123 [6], used in our test set. Shown are three consecutive frames near the start of the sequence. Note: the ground truths have been modified for qualitative evaluation (see above discussion of the UAV123 dataset).

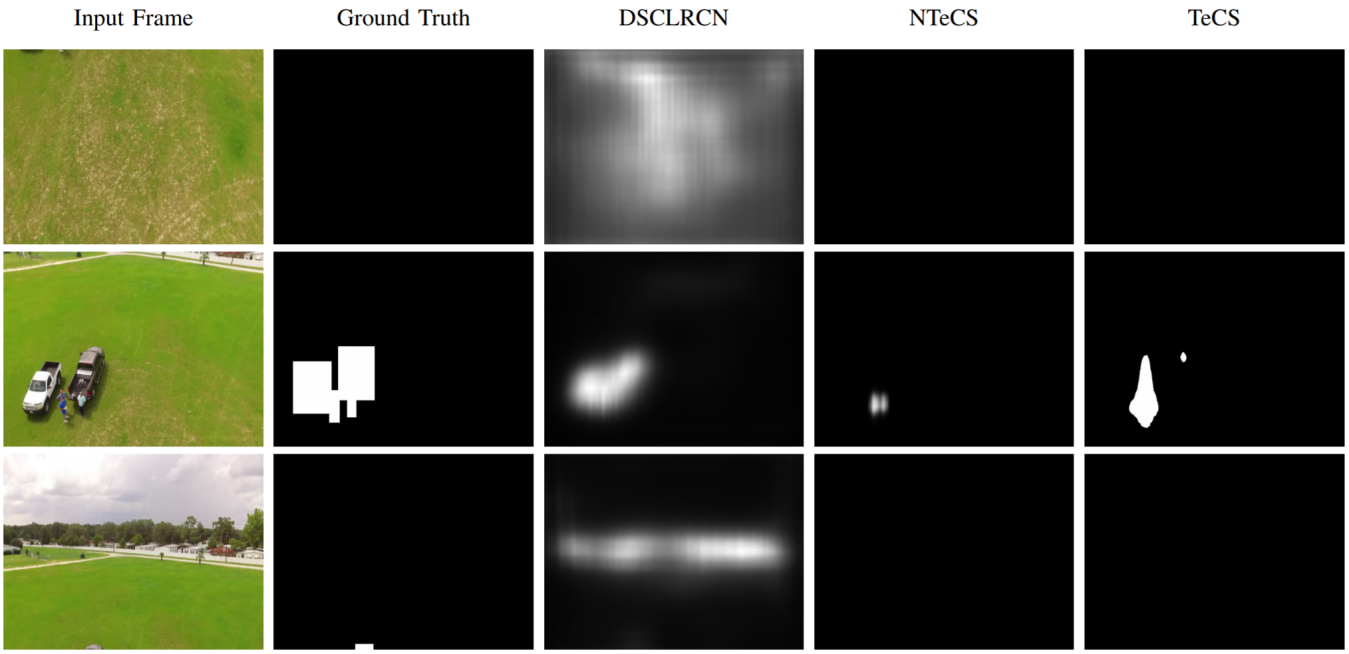

Fig. 3. Performance of baseline and proposed models on a sparse sequence. Sequence extracted from youtube.com/watch?v=V4YhIFm2no8.

model achieves improved performance when compared to the baseline DSCLRCN model. While the non-temporal NTeCS is narrowly beaten by the DSCLRCN model in NSS_{alt}, it achieves better CE_MAE and DoM scores. The temporal TeCS model achieves significantly better performance than both of these models with respect to nearly all metrics. This quantitative result is further supported by qualitative analysis.

Fig. 2 presents a qualitative comparison of the three models on a sequence from the UAV123 dataset. The temporal model outperforms both the other models, in terms of accuracy as well as consistency. The baseline model suffers from overdetection, erroneously detecting a salient object in the left half of the image in all three frames. The non-temporal model performs better than the baseline, correctly detecting both salient objects in all frames, but produces temporally inconsistent output. Both the size and the confidence of the leftmost detection varies from frame to frame, and the first and third frame have gaps within the saliency prediction of the left object. This suggests that the inclusion of temporal reasoning improves both the accuracy and consistency of the saliency prediction of the TeCS model. As the determination of saliency in a frame is based on both features in the current frame and features from past frames, any small variation in the appearance of an object that may occur frame-to-frame will produce a smaller change in the prediction, leading to more consistent output.

Fig. 3 shows another qualitative comparison of the three models on a typical UAV video. The shown frames are 20 frames apart and are taken from late in the video, $\sim$1000 frames in, with nearly all previous frames containing no salient objects. As in the previous example, the baseline DSCLRCN performs worse than the two TeCS models. This model produces extreme erroneous detections in the first and third frame where no salient object is present or is very small near the edge of the image, the reasons for which were discussed previously. The non-temporal TeCS model correctly detects no salient object in the first frame, and although it fails to detect the object in the third frame, unlike the baseline it does not produce any incorrect detections. However, in the second frame it performs worse than the baseline model, only producing a small detection near the people in the image. The temporal model performs equally in the first and third frames, but performs significantly better than the NTeCS model in the second. Despite the sudden appearance of salient objects in the sequence after a long period without any the temporal model correctly detects the salient objects, and produces no erroneous detection once the objects leave the frame.

V. Conclusion

In this work we present novel additions to the baseline DSCLRCN architecture proposed by [5], and explore the use of temporal vs. non-temporal reasoning in the form of a convLSTM layer. We present quantitative results on the UAV123 dataset [6], and qualitative results on two exemplary UAV video sequences. Our proposed TeCS model significantly outperforms the baseline DSCLRCN model.

The inclusion of temporal reasoning drastically improves the performance of the TeCS model, both in terms of accuracy, evidenced by the quantitative results, and in terms of temporal consistency, showcased in the qualitative examples. Both the quantitative and qualitative results demonstrate the importance of temporal reasoning for the task of salient object detection in UAV video, and this is likely to be a vital area to consider for future work on this topic.

REFERENCES

[1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[2] J. Sokalski, T. Breckon, and I. Cowling, "Automatic Salient Object Detection in UAV Imagery," *Proc. 25th International Conference on Unmanned Air Vehicle Systems*, pp. 11.1–11.12, 2010.

[3] Y. Zhang, A. Su, X. Zhu, X. Zhang, and Y. Shang, "Salient Object Detection Approach in UAV Video," in *Proc. SPIE Automatic Target Recognition and Navigation*, vol. 8918, 2013, p. 89180Y.

[4] S. Gotovac, V. Papić, and Ž. Marušić, "Analysis of saliency object detection algorithms for search and rescue operations," in *Proc. International Conference on Software, Telecommunications and Computer Networks*, 2016, pp. 1–6.

[5] N. Liu and J. Han, "A Deep Spatial Contextual Long-term Recurrent Convolutional Network for Saliency Detection," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3264–3274, 2018.

[6] M. Mueller, N. Smith, and B. Ghanem, "A Benchmark and Simulator for UAV Tracking," in *Proc. European Conference on Computer Vision*, 2016, pp. 445–461.

[7] C. Wang and B. Yang, "Saliency-guided object proposal for refined salient region detection," *Proc. Visual Communications and Image Processing*, pp. 1–4, 2016.

[8] Y. Zhang, X. Wang, X. Xie, and Y. Li, "Salient Object Detection via Recursive Sparse Representation," *Remote Sensing*, vol. 10, no. 4, p. 652, 2018.

[9] L. Wang, J. Xue, N. Zheng, and G. Hua, "Automatic salient object extraction with contextual cue," in *Proc. International Conference on Computer Vision*, 2011, pp. 105–112.

[10] Z. Luo, A. K. Mishra, A. Achkar, J. A. Eichel, S. Li, and P.-M. Jodoin, "Non-local Deep Features for Salient Object Detection," *Proc. Computer Vision and Pattern Recognition*, pp. 6593–6601, 2017.

[11] D. Božić-Štulić, Ž. Marušić, and S. Gotovac, "Deep Learning Approach in Aerial Imagery for Supporting Land Search and Rescue Missions," *International Journal of Computer Vision*, pp. 1–23, 2019.

[12] N. Imamoglu, W. Lin, and Y. Fang, "A Saliency Detection Model Using Low-Level Features Based on Wavelet Transform," *IEEE Transactions on Multimedia*, vol. 15, no. 1, pp. 96–105, 2013.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[14] A. Azaza and A. Douik, "Deep saliency features for video saliency prediction," in *Proc. International Conference on Advanced Systems and Electric Technologies*, 2018, pp. 355–359.

[15] H. Song, W. Wang, S. Zhao, J. Shen, and K.-M. Lam, "Pyramid Dilated Deeper ConvLSTM for Video Salient Object Detection," in *Proc. European Conference in Computer Vision*. Springer, 2018, pp. 744–760.

[16] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting," in *Proc. Advances in Neural Information Processing Systems*, 2015, pp. 802–810.

[17] A.-F. Perrin, L. Zhang, and O. Le Meur, "How well current saliency prediction models perform on uavs videos?" in *Proc. International Conference on Computer Analysis of Images and Patterns*. Springer, 2019, pp. 311–323.

[18] R. J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vision Research*, vol. 45, no. 18, pp. 2397–2416, 2005.

[19] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What Do Different Evaluation Metrics Tell Us about Saliency Models?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 3, pp. 740–757, 2019.

[20] "A data-driven metric for comprehensive evaluation of saliency models," in *Proc. International Conference on Computer Vision*. IEEE, 2015, pp. 190–198.

[21] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *CoRR*, vol. abs/1412.6980, 2015.

[22] V. Krassanakis, M. Perreira Da Silva, and V. Ricordel, "Monitoring Human Visual Behavior during the Observation of Unmanned Aerial Vehicles Videos," *Drones*, vol. 2, no. 4, 2018.