



AI-based optical-thermal video data fusion for near real-time blade segmentation in normal wind turbine operation

Jia, Xiaodong; Chen, Xiao

Published in:
Engineering Applications of Artificial Intelligence

Link to article, DOI:
[10.1016/j.engappai.2023.107325](https://doi.org/10.1016/j.engappai.2023.107325)

Publication date:
2024

Document Version
Publisher's PDF, also known as Version of record

[Link back to DTU Orbit](#)

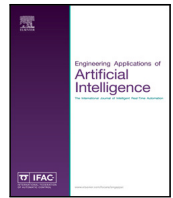
Citation (APA):
Jia, X., & Chen, X. (2024). AI-based optical-thermal video data fusion for near real-time blade segmentation in normal wind turbine operation. *Engineering Applications of Artificial Intelligence*, 127, Article 107325. <https://doi.org/10.1016/j.engappai.2023.107325>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.



AI-based optical-thermal video data fusion for near real-time blade segmentation in normal wind turbine operation

Xiaodong Jia, Xiao Chen^{*}

Technical University of Denmark, Department of Wind and Energy Systems, Roskilde, 4000, Denmark

ARTICLE INFO

Keywords:

Damage inspection
Blade segmentation
Thermal imaging
Data fusion
Multimodal complementarity
Deep learning

ABSTRACT

Blade damage inspection without stopping the normal operation of wind turbines has significant economic value. Blade segmentation is a fundamental task for blade damage inspection in the field without stopping wind turbines. This study proposes an AI-based method AQUADA-Seg to segment the images of blades from complex backgrounds by fusing optical and thermal videos taken from normal operating wind turbines. The method follows an encoder–decoder architecture and uses both optical and thermal videos to overcome the challenges associated with field application. A memory is designed between the encoder and decoder to improve the method's performance by utilizing time history information in the videos to achieve temporal complementarity. The designed memory shares information between optical and thermal modalities to achieve multimodal complementarity. We collected a large-scale dataset, i.e., 100 video pairs and over 55,000 images, of optical-thermal videos of blades in operational wind turbines to train and test the method. Experimental results show that AQUADA-Seg: i) achieves near real-time thermal-optical blade video segmentation and can analyze videos with complex backgrounds in real-world field applications; ii) achieves 0.996 and 0.981 MIoU on optical and thermal videos, respectively, outperforming state-of-the-art methods, particularly in the videos with complex backgrounds. This study provides an essential step towards automated blade damage detection using computer vision without stopping the normal operation of wind turbines.

1. Introduction

Rotor blades are critical components of wind turbine systems and often operate in harsh environments. This leads to blade failures becoming the most important contributor to wind turbine failures, followed by control system failures and electrical failures (Pérez et al., 2013; Van Bussel and Zaaier, 2001). Therefore, inspecting blades regularly to prevent blade failure is an important task in wind turbine operation and maintenance.

Blade damage inspection has seen dramatic advancement in the last decades. Traditional blade inspection method requires professionals manually check blades with rope and basket, which is labor-intensive and time-consuming. Moreover, wind turbines must be stopped when inspecting, resulting in extra turbine downtime and operative costs. To reduce inspection costs, more and more computer-vision-based methods emerged, including ground-based telescopes (Wallace and Dawson, 2009), drone-based cameras (Shihavuddin et al., 2019; Wang et al., 2019), and infrared thermography (Chen et al., 2021; Sheiati and Chen, 2023; Chen et al., 2023, 2022). For example, Shihavuddin et al. (2019) proposed a faster R-CNN and Inception-ResNet-v2 based method

that detects blade surface damages from images taken by drone-based optical cameras. Unlike optical camera based methods that only detect surface damages, some laboratory studies demonstrate that infrared thermography can detect and evaluate underneath blade damages, which are often more severe and require more attention (Chen et al., 2021; Sheiati and Chen, 2023; Chen et al., 2023, 2022). For example, Chen et al. (2023) presented a computer vision and thermal imagery based blade damage inspection method named AQUADA PLUS. This method can automatically localize, track, and evaluate multiple blade damages in blades under cyclic loading simulating operational fatigue loads. Although these methods demonstrated encouraging results in laboratories, they are too difficult to focus on blades in the field because of distraction from noisy and complex backgrounds, resulting in their severe performance degradation. Thus, blade segmentation becomes a fundamental task when applying computer-vision-based blade damage inspection methods in the field.

In the past few years, much effort has been devoted to building wind turbine blade segmentation models. For example, Xu et al. (2019) presented an optical blade segmentation method based on Canny edge

^{*} Corresponding author.

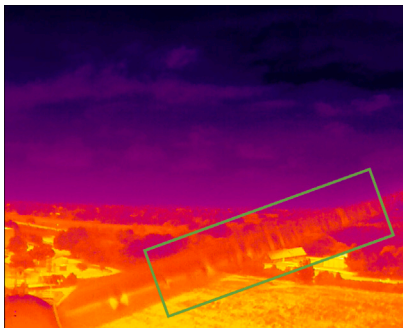
E-mail address: xiac@dtu.dk (X. Chen).



(a) A case where optical modality fails to segment the blade because the blade and cloud are mixed together, which is indicated with a green box.



(b) Complementarity from thermal modality (green box) can be utilized to improve the segmentation of case (a).



(c) A case where thermal modality fails to segment the blade because the blade and backgrounds are mixed together, which is indicated with a green box.



(d) Complementarity from optical modality (green box) can be utilized to improve the segmentation of case (c).

Fig. 1. Cases where single-modal fails to segment the blades but multimodal complementarity can be utilized to improve the segmentation performance. Thus, this study proposes using both optical and thermal modalities in blade segmentation instead of a single one.

detection and morphology. This method first segments blade images with Canny edge detection then applies morphological opening to erode and deflate segmentation masks. Tang et al. (2021) proposed a hough line detection and Otsu threshold segmentation based adaptive wind turbine blades segmentation method. This method first preliminarily locates edges of the blade line using hough line detection, then uses the grab-cut algorithm of Otsu threshold segmentation and morphological operations to segment blade images in the target area. Inspired by the huge success of deep learning (LeCun et al., 2015; Silver et al., 2016; Senior et al., 2020; Bi et al., 2023) in semantic segmentation (Oh et al., 2019; Caelles et al., 2017; Long et al., 2015; Noh et al., 2015; Ronneberger et al., 2015; Strudel et al., 2021), several deep learning based blade segmentation methods have emerged recently (Wang et al., 2022; Yang et al., 2021; Yu et al., 2023; Pérez-Gonzalo et al., 2023). For example, Yu et al. (2023) presented a U-net based thermal blade image segmentation model, in which hierarchical-split depth-wise separable convolution block is designed to obtain a balance between speed and accuracy. Wang et al. (2022) proposed a U-net based optical wind turbine segmentation model, in which two types of attention mechanisms—ECA-Net and PSA-Net—were incorporated to enhance the model's details capture ability. Yang et al. (2021) presented a blade segmentation method based on CNN and Otsu threshold. In addition, ensemble learning was introduced to improve the segmentation performance. Pérez-Gonzalo et al. (2023) proposed a U-Net and hole filling based optical blade image segmentation method. This method first employs a U-Net to generate a preliminary blade segmentation, then applies three hole filling based postprocessing steps and random forest to improve its segmentation accuracy.

1.1. Motivation

Motivation for using both optical and thermal modalities:

For segmentation: Existing blade segmentation methods either use optical or thermal modality, but in real-world applications, we found numerous cases where single-modal methods fail. Take Fig. 1(a) as an example, optical modality fails to segment the blade because the boundaries between the blade and clouds are too difficult to identify. But if a model takes both optical and thermal modalities as input, it can solve this case by utilizing complementary information provided by the thermal modality. Similarly, thermal modality fails to segment the blade, but complementarity from optical modality can be utilized to help with solving this case. Thus, we should fuse optical and thermal modalities in blade segmentation to achieve multimodal complementarity.

For damage detection: Although we focus on blade segmentation here, our long-term objective of the future study is to detect blade damage. For damage detection, the motivation for using multimodal data is twofold. On the one hand, using both optical and thermal modalities can detect surface and underneath damages simultaneously. Optical modality can be used to detect surface damage, but cannot be used to detect underneath damage, which is much more important than surface damage in wind turbine blades. Meanwhile, thermal modality can help to detect underneath damage. On the other hand, infrared thermography suffers from reflectivity-emissivity issues, which cause temperature measurement errors (Moradi and Sfarra, 2021; Gao and Tian, 2018). It has been verified that optical modality can complement thermal modality to correct reflectivity-emissivity problems (Moradi

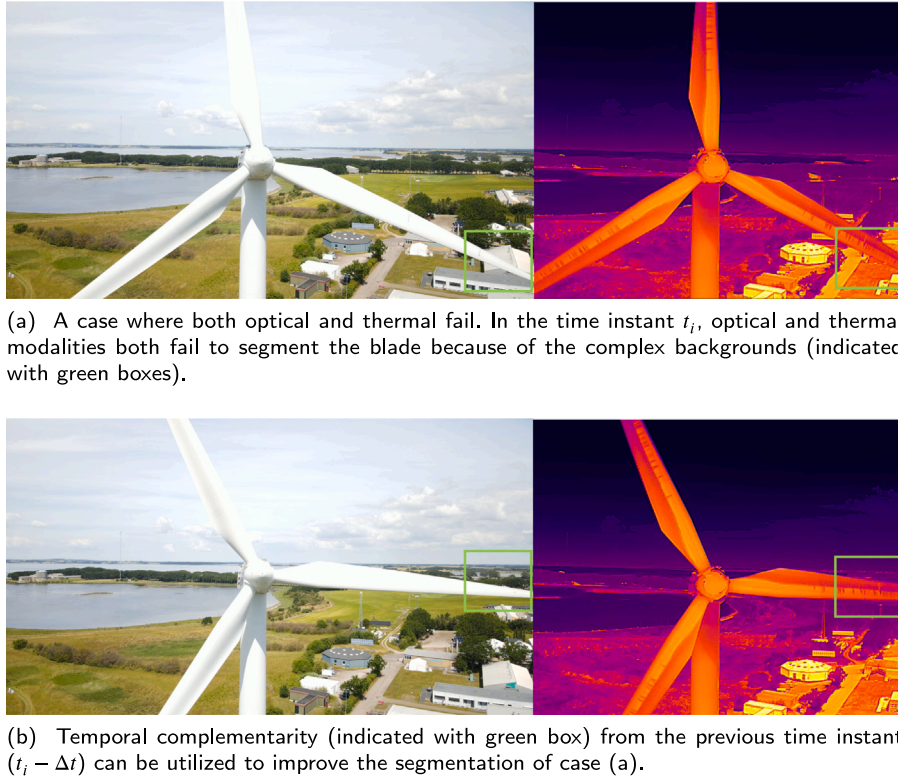


Fig. 2. Cases where optical and thermal both fail to segment the blade tip but temporal complementarity can be utilized to improve the segmentation. Thus, this study proposes using videos instead of images in blade segmentation.

et al., 2022; Tong et al., 2023). Thus, to facilitate damage detection and correcting reflectivity-emissivity issues in the future, optical and thermal modalities should be used together.

Motivation for using videos instead of images:

For segmentation: Existing blade segmentation methods only work on static images, but in real-world applications, we found many cases where optical and thermal both fail if using only images. In the cases of complex backgrounds, optical and thermal modalities may fail at the same time instant, which cannot be handled with multimodal complementarity (see Fig. 2(a)). However, blade segmentation has its unique advantages: Except for orientation, the segmentation shapes of blades do not change much at different times. If taking videos as input, a model can solve these cases by utilizing temporal complementarity in the video. Take Fig. 2 as an example, a model can utilize history complementary segmentation information from a few seconds ago (Fig. 2(b)) to help with segmenting the current frame (Fig. 2(a)). Therefore, we should use videos that contain historical information to achieve temporal complementarity.

For damage detection: Another motivation for using videos is that temporal information plays a key role in thermal-modality-based blade underneath damage detection. Because it takes time for thermal waves to reach the surface from subsurface defects, temporal information is significant in thermal-modality-based damage detection. Static images cannot provide temporal information. Therefore, not images but videos should be used.

Altogether, the objective of this work is developing a novel AI-based model, which achieves multimodal and temporal complementarity by fusing optical and thermal data, to segment blades from complex backgrounds in real-world field application videos. However, real-world hardware differences lay a challenge on our way to achieve this objective—where to get complementary information? Knowing where to get complementary information is a prerequisite for the model to utilize complementarity. Ideally, optical and thermal videos should be perfectly synchronized and have the same field of view (FOV). In

this way, when a modal fails in a certain area at a certain moment, the model can directly obtain complementary information in the corresponding area and moment from the other modality. Nevertheless, real-world optical and thermal cameras are not perfectly synchronized and they have different FOVs, spatial resolutions, frame frequencies, and reception fields (see Fig. 3). The model cannot easily get complementary information as in the ideal case. Hence, where to get complementary information is a challenge the model needs to overcome while utilizing complementarity.

1.2. Contributions

This paper contributes existing knowledge base as follows:

- This study presents a novel AI-based optical-thermal blade video segmentation model named AQUADA-Seg. AQUADA-Seg achieves near real-time optical-thermal blade video segmentation without stopping turbines and outperforms state-of-the-art blade segmentation methods.
- By taking both optical and thermal videos as input to achieve multimodal and temporal complementarity with a tailored memory, AQUADA-Seg shows that using multimodal videos instead of single-modal images significantly improves blade segmentation performance, especially in real-world applications with complex backgrounds.
- This study contributes a large-scale optical-thermal wind turbine blade video dataset. It contains 100 optical-thermal video pairs and over 55,000 images, among which 36 video pairs and 20,778 images were published to facilitate future studies.

1.3. Paper structure

The rest of this paper is organized as follows: we will start by introducing our proposed method AQUADA-Seg in Section 2, then

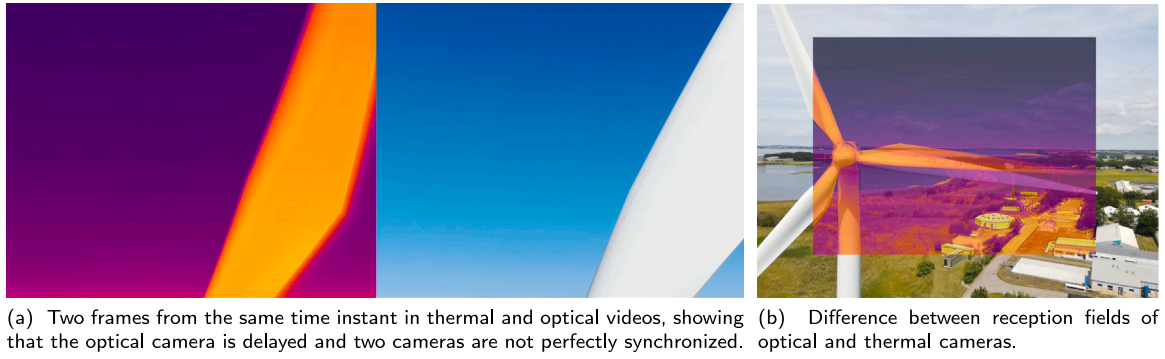


Fig. 3. Hardware differences between real-world optical and thermal cameras. (a) Thermal and optical videos are not perfectly synchronized. (b) Thermal and optical cameras have different specifications. Optical camera: pixels = 1920×1080 , FOV = 66.6° , dpi = 300; Thermal camera: pixels = 640×512 , FOV = 40.6° , dpi = 72.

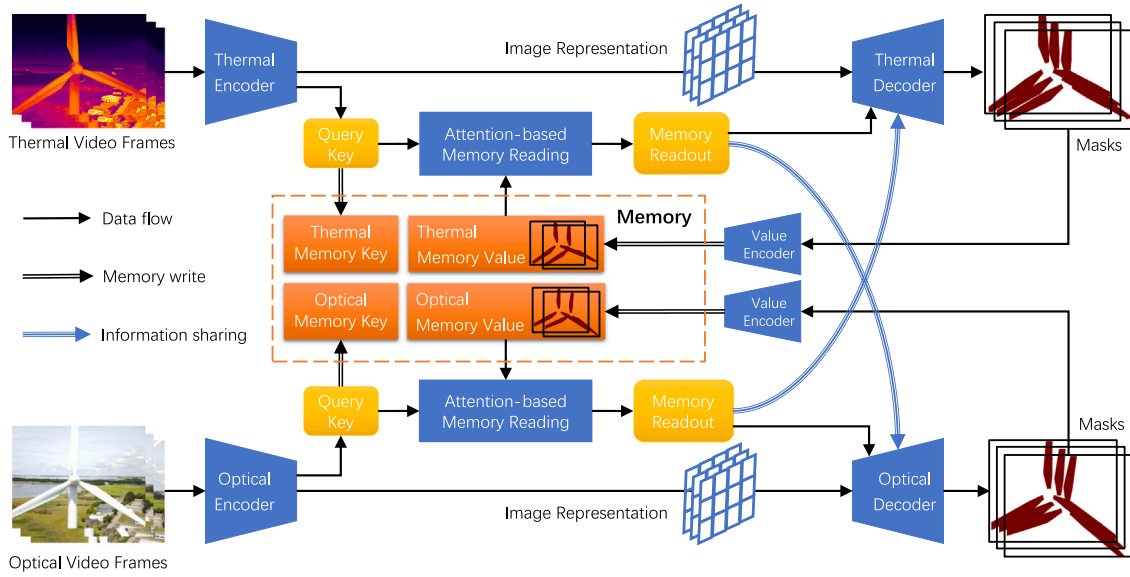


Fig. 4. The overall encoder-decoder architecture of AQUADA-Seg.

move to the experimental results including comparison with state-of-the-art and ablation studies in Section 3. Finally, we conclude the paper in Section 4.

2. Proposed method

Fig. 4 illustrates the overall architecture of AQUADA-Seg. AQUADA-Seg follows an encoder-decoder architecture. Optical and thermal modalities have their own encoder, decoder, and value encoder. For each modality, we add a memory between its encoder and decoder to store history segmentation masks. The memory adopts a key-value structure and accesses data through attention mechanism. The decoder of each modality gets input from its encoder, its memory, and importantly the other memory, then outputs a segmentation mask of the current frame. Finally, the value encoder of each modality encodes the segmentation mask and stores it in memory. In the following subsections, we will introduce the encoder-decoder architecture of AQUADA-Seg, details of the designed memory, the loss function, and our collected optical-thermal wind turbine blade video dataset respectively.

2.1. Encoder-decoder architecture

Overall, AQUADA-Seg is an encoder-decoder style network. Each modality has its own encoder, decoder, and lightweight value encoder.

We built the encoders and decoder following segmentation network STCN (Cheng et al., 2021). Specifically:

Encoder of each modality takes an image as input and outputs a representation of the image and a query key. The representation is the “code” and will be input into the decoder. The query key, which also works as a memory key, will be used when reading memory. Following common practice (Cheng et al., 2021; Oh et al., 2019), we constructed the encoder based on Resnet-50 (He et al., 2016), removing its last convolutional layer and classification layer.

Decoder of each modality outputs the segmentation mask of the current input image. It takes the following three types of information as inputs:

- Representation of the input image, which is obtained from the encoder.
- History segmentation information read from memory.
- Multimodal complementary information obtained from the counterpart modality.

We constructed the decoder following STM network (Oh et al., 2019). In particular, decoder first fuses the image representation and memory readout, which get from encoder and memory respectively, with a group convolutional neural network. Then it upscales the fused feature. Finally, masks outputted by the decoder will be bilinearly upsampled to the original resolution.

Value encoder of each modality encodes the information that will be stored in the value part of memory. Since the memory of each

modality stores the history segmentation masks, value encoder encodes the masks generated by the decoder. Because segmentation masks are easier to encode than input images, we construct the value encoder based on a lightweight network—Resnet-18 (He et al., 2016), removing its last convolutional layer and classification layer.

2.2. Memory

On top of encoder–decoder architecture, we design a memory to utilize temporal complementarity and multimodal complementarity to enhance the model's performance. In the following subsections, we first introduce the details of this memory, including key–value memory structure, memory writing, attention-based memory reading, and memory management, then move on to how AQUADA-Seg utilizes temporal and multimodal complementarity with this memory.

2.2.1. Memory details

Key–value Memory Structure

As illustrated in Fig. 4, we designed a key–value memory for optical and thermal modalities respectively. Key works as indexes, responsible for memory reading. Value stores history segmentation masks. Key comes from the encoder, which is essentially a compressed image representation. Value comes from the value encoder, which is essentially a compressed segmentation mask. After decoder outputs a segmentation mask of the current frame, the model updates memory by adding a new key and value to it.

Attention-based Memory Reading

AQUADA-Seg reads memory in an attention-based way. When segmenting the $(N + 1)$ th frame of input video, the model first encodes it with encoder, then starts to read memory. At this time, memory stores segmentation information of previous N frames. Let $\mathbf{k}^M \in \mathbb{R}^{D^k \times N \times H \times W}$ be the memory key, $\mathbf{v}^M \in \mathbb{R}^{D^v \times N \times H \times W}$ be the memory value, $\mathbf{k}^Q \in \mathbb{R}^{D^k \times N \times H \times W}$ be the query key obtained from the encoder, where D^k and D^v denote the dimensions of key and value, H and W denote the spatial dimensions. We employ a similarity function $s(\cdot)$ to compute the similarity matrix $\mathbf{S} \in \mathbb{R}^{N \times H \times W \times H \times W}$ of \mathbf{k}^M and \mathbf{k}^Q . This process can be written as:

$$\mathbf{S}_{ij} = s(\mathbf{k}_i^M, \mathbf{k}_j^Q). \quad (1)$$

In practice, we use the L2 similarity function proposed in STCN (Cheng et al., 2021), and normalize the similarity matrix with $\sqrt{D^k}$. Then, we let \mathbf{S} pass a softmax function to get the softmax-normalized attention weight matrix $\mathbf{W} \in \mathbb{R}^{N \times H \times W \times H \times W}$, which can be represented by:

$$\mathbf{W}_{ij} = \frac{\exp(\mathbf{S}_{ij})}{\sum_n (\exp(\mathbf{S}_{nj}))}. \quad (2)$$

Finally, the memory readout of the $(N + 1)$ th frame \mathbf{m}_{N+1}^Q can be computed as the weighted sum of memory value, which can be written as:

$$\mathbf{m}_{N+1}^Q = \mathbf{v}^M \mathbf{W}. \quad (3)$$

The memory readout \mathbf{m}_{N+1}^Q works as the temporal complementary information and will be input into the decoder to assist the segmentation of the $(N + 1)$ th frame.

Memory Management

As we update memory for each frame of the video input, the memory size gradually increases as the number of frames increases. If we do not manage memory, it will explode soon. Especially when training the model with long videos. Following Cheng and Schwing (2022), we divide memory into different segments and start to clean up the oldest saved masks when the memory reaches its limit. In addition, since wind turbines rotate periodically, blade segmentation does not require large memory.

2.2.2. Memory design for utilizing temporal complementarity and multimodal complementarity

With the designed memory, AQUADA-Seg can utilize temporal complementarity when segmenting new frames. Specifically, AQUADA-Seg saves history masks in memory and reads these masks when segmenting the current frame. Because the shape of blade segmentations does not change much at different times, historical segmentation information is of great value for segmenting the current frame. Moreover, attention-based memory reading helps AQUADA-Seg find the most useful information for segmenting the current frame. Thus, AQUADA-Seg utilizes temporal complementarity when segmenting new frames.

AQUADA-Seg utilizes multimodal complementarity by sharing complementary information between optical and thermal modalities via the memory. As described above, AQUADA-Seg inputs the information read from memory into the decoder of the current modality to help with its segmentation. Inspired by Jia et al. (2023), we made AQUADA-Seg also share this information with the other modality (see blue lines in Fig. 4). Since optical and thermal videos are almost synchronized, information read from optical memory, which is most useful for segmenting the current optical frame, is also of great help for segmenting the current thermal frame, and vice versa. Thus, when a modality fails – its encoder and memory fail to provide useful information for its segmentation – it still can utilize complementary information obtained from the other modality to assist its segmentation. With this cross-modal complementary information sharing, AQUADA-Seg utilizes multimodal complementarity in blade segmentation.

2.3. Loss function

Following previous semantic segmentation studies (Cheng et al., 2021; Cheng and Schwing, 2022; Wang et al., 2022), we employ binary cross entropy loss (BCE loss) and Dice loss (Milletari et al., 2016) to train AQUADA-Seg. The loss function of AQUADA-Seg can be written as:

$$Loss = L_{BCE}^{Thermal} + L_{BCE}^{Optical} + \alpha(L_{Dice}^{Thermal} + L_{Dice}^{Optical}), \quad (4)$$

$$L_{BCE}^{Thermal} = -\frac{1}{N} \sum_{i=1}^N (y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)), \quad (5)$$

$$L_{Dice}^{Thermal} = 1 - 2 \sum_{i=1}^N \frac{y_i \cdot \hat{y}_i}{y_i + \hat{y}_i}, \quad (6)$$

where α is a trade-off parameter, N is the number pixels in a thermal frame, y_i is the binary label of the i th pixel from this frame, \hat{y}_i is the model's prediction of the same pixel. Since $L_{BCE}^{Optical}$ is similar to $L_{BCE}^{Thermal}$ and $L_{Dice}^{Optical}$ is similar to $L_{Dice}^{Thermal}$, we do not repeat them here.

2.4. Optical-thermal wind turbine blade video dataset

To train AQUADA-Seg, we collected a large-scale optical-thermal wind turbine blade video dataset. Moreover, we make it publicly available to facilitate future studies. It can be accessed here.¹ This dataset contains 100 optical-thermal video pairs and over 55,000 images from 22 different wind turbines. The videos are collected from each turbine at different time and under different environmental conditions. We only published the data collected from DTU Vestas V52 wind turbine, i.e., 36 optical-thermal video pairs and 20,778 images. The data from other 21 commercial turbines is not published due to confidentiality. Table 1 tabulates the information of this dataset. Table 2 compares some existing datasets (Zampokas et al., 2022; Pérez-Gonzalo et al., 2023; Wang et al., 2022; Yu et al., 2023) that can be used for blade segmentation. To the best of our knowledge, our dataset is the largest wind turbine blade dataset to date.

¹ <https://aquada-go.github.io/>.

Table 1

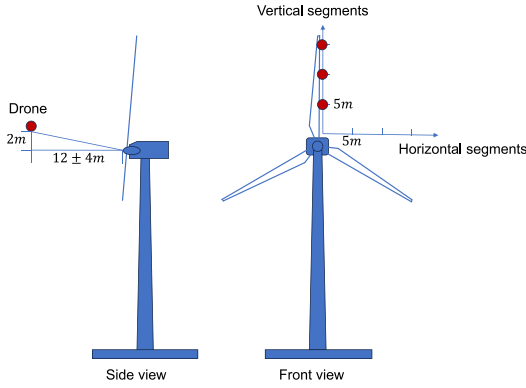
Information of the optical-thermal wind turbine blade video dataset used in this study.

Number of optical-thermal videos (in pairs)	100
Number of images	55 880
Number of wind turbines	22 ^a
Train/test size	70/30
Turbine IDs in train set	(1, 2, 3, ..., 17, 18)
Turbine IDs in test set	(1, 2, 5, 6, 9, 13, 15, 16, 19, 20, 21, 22)
Average frames in each video	279
Optical video frame size in pixels	1920 × 1080
Thermal video frame size in pixels	640 × 512

^a We collect videos from each turbine at different time and under different environmental conditions.**Table 2**

Comparison of existing datasets that can be used for wind turbine blade segmentation.

Reference	Number of images	Optical	Thermal	Publicly available
Wang et al. (2022)	330	✓		No
Yu et al. (2023)	312		✓	No
Zampokas et al. (2022)	224	✓		Yes
Pérez-Gonzalo et al. (2023)	2032	✓		No
This study	55 880 ^a	✓	✓	Yes

^a Due to confidentiality, only a part of the data was published, i.e., 36 optical-thermal video pairs and 20,778 images. It can be found at <https://aquada-go.github.io/>.**Fig. 5.** Drone-based optical-thermal blade video data acquisition when the wind turbine is in normal operation.

All blade videos are taken with DJI Zenmuse H20T² or DJI Mavic 2 Enterprise Advanced³ while wind turbines are in normal operation. We fixed the frame frequencies of the optical and thermal cameras to 30 FPS. The fusion color palette is chosen for thermal cameras. We first fly the drone to a position where the horizontal distance from the hub nose is 12 ± 4 m and the vertical distance is 2 m (see Fig. 5). We tilt up the camera 15 degrees to avoid taking videos of the thermal source from the nacelle. Then, we take both optical and thermal videos in pairs. For long blades, we take videos of them horizontally or vertically in several segments. The interval between videoing positions of different segments is about 5 m. We take videos from both sides of the blades, i.e., both from upwind and downwind directions. In addition, to increase the diversity of data and improve the robustness of the model, we also take various videos from different angles and distances. Fig. 6 demonstrates some optical-thermal images in this dataset and their segmentation masks.

3. Experimental results

In this section, we first compare AQUADA-Seg with state-of-the-art methods, then conduct ablation studies to evaluate the effectiveness of

our two major differences, i.e., multimodal vs. single-modal and videos vs. images.

3.1. Metrics

We use two commonly used segmentation metrics, i.e., MIoU and MPA to compare all the results. Mean Intersection over Union (MIoU) is a common metric for semantic segmentation. It computes the coincidence ratio between ground truth and the model's prediction. MIoU is defined as follows:

$$\text{MIoU} = \frac{1}{M} \sum_{i=1}^M \left(\frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \right), \quad (7)$$

where M is the number of classes, TP is true positive, FP is false positive, FN is false negative, and FP is false positive. Mean Pixel Accuracy (MPA) also is a popular segmentation metric, which computes the mean of the right predicted pixel ratio from different classes. MPA is defined as follows:

$$\text{MPA} = \frac{1}{M} \sum_{i=1}^M \left(\frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \right). \quad (8)$$

3.2. Comparison with state-of-the-art

3.2.1. Settings

Compared Methods: To the best of our knowledge, we are the first to segment wind turbine blades with videos and there are not any wind turbine blade video segmentation methods. Hence, we compare AQUADA-Seg with two state-of-the-art blade image segmentation methods—Improved-UNet-Thermal (IUNet-T) (Yu et al., 2023) and Improved-UNet-Optical (IUNet-O) (Wang et al., 2022). Table 3 gives an overview of these two state-of-the-art methods and AQUADA-Seg.

Because these methods work either on optical or thermal data, we train them with data only from one modality. Because these methods only work on images, we test them on all video frames and take the average as their results on video. According to the study (Wang et al., 2022), 10% of training data is used as the validation set for IUNet-O.

Data Preprocessing: AQUADA-Seg shares complementary information between different modalities. To unify the shape of shared information across different modalities as well as to reduce computation burden, we first resize frames of optical and thermal videos to 852×480 . Then, we conduct data augmentation, including random rotation, random crop, random horizontal flip, and random color jitter.

² <https://enterprise.dji.com/zenmuse-h20-series>.³ <https://enterprise.dji.com/mavic-2-enterprise-advanced>.

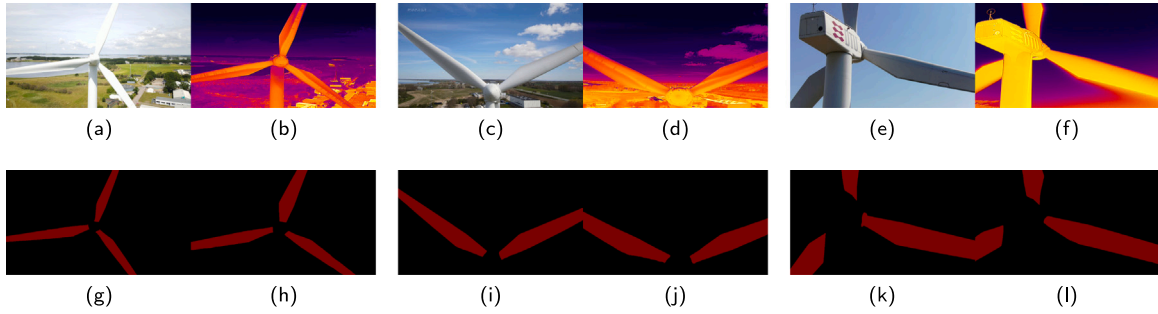


Fig. 6. Some optical-thermal images in our dataset and their segmentation masks. We can see these images vary considerably in the videoing distance, background, videoing angle, and lighting, indicating they are close to the images taken in real-world applications.

Table 3

Overview of AQUADA-Seg and two relevant state-of-the-art methods that only work on single-modal data.

Methods	Working modality	Backbone	Proposed year	Reference
IUNet-T	Thermal	UNet	2023	Yu et al. (2023)
IUNet-O	Optical	Res-UNet	2022	Wang et al. (2022)
AQUADA-Seg	Optical and thermal	Resnet and Encoder-decoder	2023	This study

Table 4

Comparison of MIoU between different methods (higher is better, N.A. for not applicable).

	Methods		
	IUNet-T (Yu et al., 2023)	IUNet-O (Wang et al., 2022)	AQUADA-Seg (This study)
Thermal	0.916	N.A.	0.981
Optical	N.A.	0.919	0.996

Table 5

Comparison of MPA between different methods (higher is better, N.A. for not applicable).

	Methods		
	IUNet-T (Yu et al., 2023)	IUNet-O (Wang et al., 2022)	AQUADA-Seg (This study)
Thermal	0.953	N.A.	0.992
Optical	N.A.	0.969	0.998

Training Details: Following previous work (Cheng et al., 2021; Cheng and Schwing, 2022), we train AQUADA-Seg with different stages. In the first stage, we train the model with static images. In the second stage, we mixed videos from DAVIS video segmentation dataset (Perazzi et al., 2016; Pont-Tuset et al., 2017) and our collected dataset to train the model. At this stage, since DAVIS dataset only contains optical data, we make a copy of the optical data as thermal data. In the third stage, we train the model with our collected optical-thermal wind turbine blade video data. These three stages were iterated 5k, 8k, and 8k times respectively. The model is implemented with Pytorch (v1.13.0) and optimized by AdamW with a beginning learning rate of 1×10^{-5} . Besides, MultiStepLR is employed to adjust the learning rate. We train the model with a computer provided by Denmark Technical University Computing Centre. This computer is equipped with two 32-core Intel Xeon Gold 6226R CPUs, 756 GB of memory, and two NVIDIA A100 (40 GB) GPUs. The entire training takes approximately 52 h.

3.2.2. Results and discussion

In the test phase, we tested the model only on a single GPU. To reduce the impact of hardware on the results, we ran test 10 times and recorded the average.

Tables 4 and 5 report the blade segmentation results of all the methods in terms of MIoU and MPA, respectively. On thermal data, we can see that AQUADA-Seg outperforms the state-of-the-art method (IUNet-T) by 0.065 (0.981 – 0.916) or 7.096% (0.065/0.916) on MIoU and 0.039 (0.992 – 0.953) or 4.092% (0.039/0.953) on MPA. On optical data, AQUADA-Seg outperforms the state-of-the-art method (IUNet-O) by 0.077 (0.996 – 0.919) or 8.379% (0.077/0.919) on MIoU and 0.029

(0.998 – 0.969) or 2.993% (0.029/0.969) on MPA. These results demonstrate the superiority of AQUADA-Seg on both thermal and optical wind turbine blade segmentation.

Across 10 tests, the average maximum GPU memory allocated of AQUADA-Seg was 1584 MB. The average test FPS of AQUADA-Seg is 26.75, showing that it achieves near real-time wind turbine blade segmentation without stopping turbines. Notably, AQUADA-Seg segments RGB and thermal videos simultaneously. This new capability opens vast opportunities for real-world applications. For example, AQUADA-Seg provides at least the following three possibilities if it is applied to blade damage detection: (i) Unlike previous methods that either detect surface damages based on optical data or detect underneath damages based on thermal data, detecting both these damages simultaneously is possible now. (ii) Unlike previous image-based blade damage detection methods that can only obtain damage status at a certain moment. AQUADA-Seg enables the detection and intervention of blade damages in near real-time, thus avoiding significant property loss. (iii) Obtaining detailed damage progress in normal operating wind turbines is possible. By analyzing the damage progress, blade researchers not only can get a better understanding of damages but also gain clues for blade structure designs.

To intuitively compare the performance of these methods, we conducted a case study. Specifically, we selected some cases with simple or complex backgrounds from the test set and compared segmentations of these methods. Table 6 compares the results. From Table 6 we can see that:

- Both relevant methods and AQUADA-Seg are capable of segmenting simple cases. The backgrounds of these cases are simple, with

Table 6

Comparison of segmentations from AQUADA-Seg and state-of-the-art methods. All these methods are capable of handling cases where backgrounds are simple and boundaries between blades and backgrounds are clear. In cases where the background is complex and blades and background are mixed, AQUADA-Seg clearly outperforms state-of-the-art methods.

	Thermal input	Segmentation using IUNet-T Yu et al. (2023)	Segmentation using AQUADA-Seg (this study)	Optical input	Segmentation using IUNet-O Wang et al. (2022)	Segmentation using AQUADA-Seg (this study)
Cases with clear and simple background						
Cases with complex background						

only sky in background and relatively few clouds, which makes clear boundaries between blades and backgrounds.

- For complex cases, however, AQUADA-Seg clearly outperforms the relevant methods. The backgrounds of these cases are complex, with either a landscape (e.g., column 2, row 7 and column 5, row 6) or a dense layer of clouds (e.g., column 2, row 5 and column 5, row 7). Blades and backgrounds are mixed together, and the boundary between them is difficult to distinguish. Single-modal based state-of-the-art methods fail to segment these cases. However, AQUADA-Seg achieves remarkable results even in these complex cases due to multimodal complementarity.
- Our dataset contains numerous complex cases, which are closer to real-world field situations.

We designed software with a user-friendly GUI as shown in the video.⁴

3.3. Multimodal vs. single-modal

The first big difference between AQUADA-Seg and existing blade segmentation methods is that AQUADA-Seg takes multimodal data as input while existing methods take single-modal data as input. Here, we investigate the effectiveness of multimodal data on blade segmentation with experiments.

3.3.1. Settings

In this experiment, we compare the following three methods:

- This study, the AQUADA-Seg method.
- Thermal-only, which is implemented by removing optical parts from AQUADA-Seg.
- Optical-only, which is implemented by removing thermal parts from AQUADA-Seg.

For AQUADA-Seg, we use the same experimental setting as in Section 3.2. For Thermal-only and Optical-only, we also train them with 3 stages and the same iterations that are used in AQUADA-Seg. But in the third stage, we train Thermal-only only with thermal data and train Optical-only only with optical data. Other settings stay unchanged.

Table 7

Comparison of contribution from different modalities in terms of MIOU.

	Modalities		
	Thermal-only	Optical-only	Thermal and optical (This study)
Thermal	0.941	N.A.	0.981
Optical	N.A.	0.953	0.996

Table 8

Comparison of contribution from different modalities in terms of MPA.

	Modalities		
	Thermal-only	Optical-only	Thermal and optical (This study)
Thermal	0.968	N.A.	0.992
Optical	N.A.	0.980	0.998

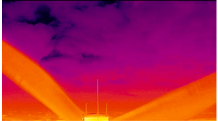




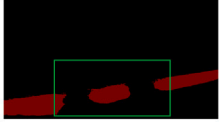


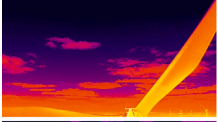
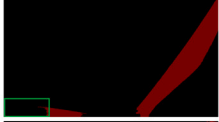



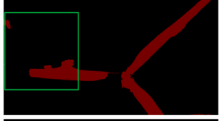

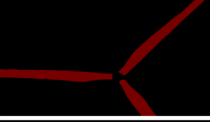




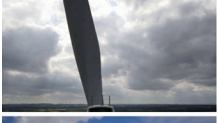
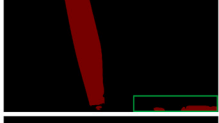
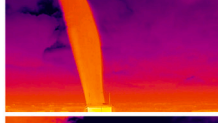


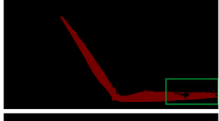

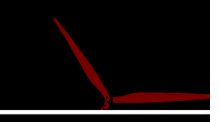

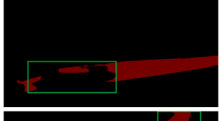
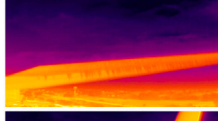



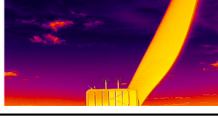

3.3.2. Results and discussion

Tables 7 and 8 compare the results of these methods in terms of MIOU and MPA. From these results, we can see that: (i) The model trained with multimodal data outperforms the model trained with single-modal data. This confirms that using multimodal data can improve the performance of blade segmentation. (ii) Among the methods trained with single-modal data, the model trained with optical data outperforms that trained with thermal data. This may be because unlike thermal data which can only provide temperature information, optical data can provide richer information, such as color information and texture information. Thermal modality is more likely to fail than optical modality. (iii) If you mainly focus on thermal blade segmentation, introducing optical modality and utilizing the complementarity of multimodal data can significantly improve the segmentation performance.

To intuitively investigate the effectiveness of multimodal complementarity, we conducted a case study. Table 9 compares the results from models trained with single-modal and multimodal data. From Table 9 we can see that: Since the information provided by a single modality is limited, it is inevitable that single-modal fails. For example case 1, case 3, and case 8. In these cases, blades and backgrounds are mixed together. It is too difficult for models trained with single-modal data to handle these cases (see the third column in Table 9). AQUADA-Seg takes both thermal and optical modalities as inputs and

⁴ <https://aquada-go.github.io/>.

Table 9
Segmentation comparison between single-modal methods and our multimodal method.

Case	Single-modal input (thermal or optical)	Segmentation from single-modal methods (Thermal-Only or Optical-Only)	Complementary-modal input (optical or thermal)	Segmentation from this study with both optical and thermal input
1				
2				
3				
4				
5				
6				
7				
8				
9				

can handle these cases by utilizing the complementarity between these two modalities. Thus, better segmentation performance is achieved.

3.4. Videos vs. images

The second big difference between AQUADA-Seg and existing blade segmentation methods is that AQUADA-Seg takes videos as input while existing methods take images as input. Here, we investigate the effectiveness of temporal complementarity on blade segmentation by comparing the results from models with access to different amounts of temporal information.

3.4.1. Settings

We investigate the effectiveness of temporal complementarity by controlling the temporal information that can be utilized by the model. AQUADA-Seg saves history segmentation information in memory and updates memory for every segmented frame. Therefore, we can control the max number of history frames that AQUADA-Seg can access by

controlling memory size. Thus simulating situations where the model obtains different amounts of temporary information. Here, we compare the performance of AQUADA-Seg with access to different numbers of frames, including 0, 25, 75, 100, 150, 200, 250, and 300. MIoU is selected as the evaluation metric. Other settings stay unchanged as in Section 3.2.

3.4.2. Results and discussion

Fig. 7 illustrates the performance of AQUADA-Seg with access to different numbers of history video frames. From Fig. 7, we can see that: when the number of frames AQUADA-Seg can access is less than 150, the performance of AQUADA-Seg gradually improves with the growth of Memory. When this number exceeds 150, the performance of the model gradually stabilizes. This verified that we can indeed improve the blade segmentation performance by utilizing temporal complementarity with a designed memory. In addition, since wind turbine blade videos are periodic, there is an upper bound to improve the segmentation performance by increasing the memory size. The recommended memory size for our case is 150 frames.

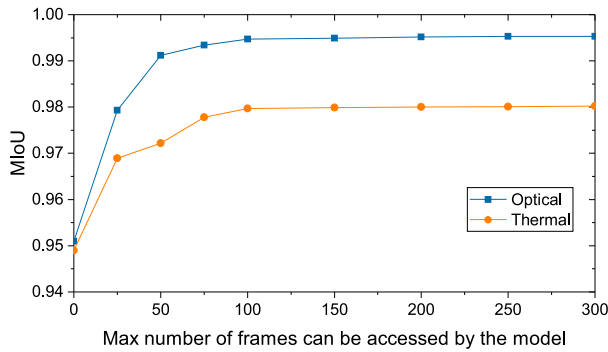


Fig. 7. Comparison of the performance of AQUADA-Seg with access to different numbers of history video frames.

3.5. AQUADA-Seg's robustness against noisy input

Although deep-learning-based methods achieve state-of-the-art performance in various real-world tasks, their robustness against input noise is still a hot topic, because noisy input may cause dramatic performance degradation, thereby leading to disasters in real-world applications (Maulik et al., 2020; Li et al., 2020). In this subsection, we investigate the robustness of AQUADA-Seg against noisy input.

3.5.1. Settings

In this experiment, we investigate the relationship between performance of AQUADA-Seg and the magnitude of input noise. We first randomly replace 1%, 2%, 3%, 4%, and 5% of optical videos with noisy input in the test set. Then observe AQUADA-Seg's performance under noisy input. We construct noisy input by weighted stacking frames from different time instants of the same video. Specifically, for a randomly selected test optical video, we first randomly select a number Δt between 3 and 10. Then, starting from Frame1, we stack Frame1 and Frame(1 + Δt) with weights of 0.8 and 0.2. Fig. 8 shows a frame of the noisy input.

3.5.2. Results and discussion

Table 10 shows the optical segmentation performance of AQUADA-Seg under different magnitudes of input noise. From Table 10 we can see that: with the increase of magnitude of input noise, AQUADA-Seg's performance drops, but slightly. This may be because when optical modality is affected by a small magnitude of noise, the model can use the information of thermal modality to assist its segmentation. Hence, we can conclude that AQUADA-Seg has high robustness when a single modality is affected by noisy input.

4. Conclusion and future work

In this paper, we propose AQUADA-Seg, an AI-based encoder-decoder style method that achieves near real-time optical-thermal wind turbine blade video segmentation. AQUADA-Seg fuses both optical and thermal videos captured from normal operating wind turbines and improves the blade segmentation performance by utilizing temporal and multimodal complementarity with a tailored memory. AQUADA-Seg utilizes temporal complementarity by storing history segmentations in the memory and reading them when segmenting new frames. In addition, AQUADA-Seg utilizes multimodal complementarity by sharing complementary segmentation information via the memory. Experimental results from a large-scale optical-thermal video dataset show that AQUADA-Seg considerably outperforms state-of-the-art optical or thermal blade segmentation methods, particularly in cases when complex backgrounds are present in real-world applications.

For neural-network-based methods, reliability of the generated results is important, especially in real-world applications where huge



Fig. 8. A frame of noisy input. We construct noisy input by weighted stacking the original frame and another frame at a different time instant from the same video.

Table 10

AQUADA-Seg's performance under different magnitudes of input noise.

Performance	Magnitudes of the input noise					
	0%	1%	2%	3%	4%	5%
MIoU	0.996	0.995	0.994	0.990	0.983	0.976
MPA	0.998	0.997	0.996	0.993	0.988	0.981

property losses and personal casualties may occur. In the future, we intend to extend AQUADA-Seg by introducing probabilistic neural network, enabling the method to output a prediction distribution rather than just the best prediction, thus, assessing the uncertainty of the method's output. Another future study is near real-time wind turbine blade damage detection by utilizing multimodal complementarity.

CRediT authorship contribution statement

Xiaodong Jia: Developed the method, Implemented the codes, Wrote the original manuscript. **Xiao Chen:** Generated research ideas, Designed and supervised the study, Reviewed and revised the manuscript and acquired the funding.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The study is supported by AQUADA-GO project: Automated blade damage detection and near real-time evaluation for operational offshore wind turbines (64022–1025), funded by The Energy Technology Development and Demonstration Programme (EUDP). We would like to thank Jesper Smit, Pjort Kat, Søren William Lund, and Mohammad Hedayatzaadeh for blade video data collection.

References

- Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., Tian, Q., 2023. Accurate medium-range global weather forecasting with 3D neural networks. *Nature* 1–6.
- Caelles, S., Maninis, K.-K., Pont-Tuset, J., Leal-Taixe, L., Cremers, D., Van Gool, L., 2017. One-shot video object segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Chen, X., Janeliukstis, R., Sarhadi, A., 2022. Thermographic data analytics-based damage characterization in a large-scale composite structure under cyclic loading. *Compos. Struct.* 290, 115525.

- Chen, X., Sheiati, S., Shihavuddin, A., 2023. AQUADA plus: Automated damage inspection of cyclic-loaded large-scale composite structures using thermal imagery and computer vision. *Compos. Struct.* 318, 117085.
- Chen, X., Shihavuddin, A., Madsen, S.H., Thomsen, K., Rasmussen, S., Branner, K., 2021. AQUADA: Automated quantification of damages in composite wind turbine blades for LCOE reduction. *Wind Energy* 24 (6), 535–548.
- Cheng, H.K., Schwing, A.G., 2022. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In: *European Conference on Computer Vision*. pp. 640–658.
- Cheng, H.K., Tai, Y.-W., Tang, C.-K., 2021. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. In: *Advances in Neural Information Processing Systems*. vol. 34, pp. 11781–11794.
- Gao, Y., Tian, G.Y., 2018. Emissivity correction using spectrum correlation of infrared and visible images. *Sensors Actuators A* 270, 8–17.
- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 770–778.
- Jia, X., Jing, X.-Y., Sun, Q., Chen, S., Du, B., Zhang, D., 2023. Human collective intelligence inspired multi-view representation learning — Enabling view communication by simulating human communication mechanism. *IEEE Trans. Pattern Anal. Mach. Intell.* 45 (6), 7412–7429.
- LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. *Nature* 521 (7553), 436–444.
- Li, X., Jia, X., Jing, X.-Y., 2020. Negative-aware training: be aware of negative samples. In: *ECAI 2020*. pp. 1269–1275.
- Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Maulik, R., Fukami, K., Ramachandra, N., Fukagata, K., Taira, K., 2020. Probabilistic neural networks for fluid flow surrogate modeling and data recovery. *Phys. Rev. Fluids* 5 (10), 104401.
- Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: *2016 Fourth International Conference on 3D Vision (3DV)*. pp. 565–571.
- Moradi, M., Ghorbani, R., Sfarra, S., Tax, D.M., Zarouchas, D., 2022. A spatiotemporal deep neural network useful for defect identification and reconstruction of artworks using infrared thermography. *Sensors* 22 (23), 9361.
- Moradi, M., Sfarra, S., 2021. Rectifying the emissivity variations problem caused by pigments in artworks inspected by infrared thermography: A simple, useful, effective, and optimized approach for the cultural heritage field. *Infrared Phys. Technol.* 115, 103718.
- Noh, H., Hong, S., Han, B., 2015. Learning deconvolution network for semantic segmentation. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- Oh, S.W., Lee, J.-Y., Xu, N., Kim, S.J., 2019. Video object segmentation using space-time memory networks. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 9226–9235.
- Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A., 2016. A benchmark dataset and evaluation methodology for video object segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 724–732.
- Pérez, J.M.P., Márquez, F.P.G., Tobías, A., Papaalias, M., 2013. Wind turbine reliability analysis. *Renew. Sustain. Energy Rev.* 23, 463–472.
- Pérez-Gonzalo, R., Espersen, A., Agudo, A., 2023. Robust wind turbine blade segmentation from RGB images in the wild. In: *IEEE International Conference on Image Processing*.
- Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L., 2017. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*.
- Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. pp. 234–241.
- Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Židek, A., Nelson, A.W., Bridgland, A., et al., 2020. Improved protein structure prediction using potentials from deep learning. *Nature* 577 (7792), 706–710.
- Sheiati, S., Chen, X., 2023. Deep learning-based fatigue damage segmentation of wind turbine blades under complex dynamic thermal backgrounds. *Struct. Health Monit.* 1.
- Shihavuddin, A., Chen, X., Fedorov, V., Nymark Christensen, A., Andre Brogaard Riis, N., Branner, K., Bjorholm Dahl, A., Reinhold Paulsen, R., 2019. Wind turbine surface damage detection by deep learning aided drone inspection analysis. *Energies* 12 (4), 676.
- Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al., 2016. Mastering the game of go with deep neural networks and tree search. *Nature* 529 (7587), 484–489.
- Strudel, R., Garcia, R., Laptev, I., Schmid, C., 2021. Segmenter: Transformer for semantic segmentation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 7262–7272.
- Tang, Z., Peng, Y., Wang, W., Cao, G., Chao, W., 2021. Adaptive segmentation method for wind turbine blades combining hough line detection and grab-cut algorithm. *J. Electron. Meas. Instrum.* 35 (4), 161–168.
- Tong, Z., Cheng, L., Xie, S., Kersemans, M., 2023. A flexible deep learning framework for thermographic inspection of composites. *NDT E Int.* 139, 102926.
- Van Bussel, G., Zaijier, M., 2001. Reliability, availability and maintenance aspects of large-scale offshore wind farms, a concepts study. In: *Proceedings of MAREC*. vol. 2001.
- Wallace, Jr., J., Dawson, M., 2009. O&m strategies: wind turbine blades. *Renew. Energy Focus* 10 (3), 36–41.
- Wang, L., Yang, J., Huang, C., Luo, X., 2022. An improved U-net model for segmenting wind turbines from UAV-taken images. *IEEE Sensors Lett.* 6 (7), 1–4.
- Wang, L., Zhang, Z., Luo, X., 2019. A two-stage data-driven approach for image-based wind turbine blade crack inspections. *IEEE/ASME Trans. Mechatronics* 24 (3), 1271–1281.
- Xu, H., Xu, X., Zuo, Y., 2019. Applying morphology to improve canny operator's image segmentation method. *J. Eng.* 2019 (23), 8816–8819.
- Yang, X., Zhang, Y., Lv, W., Wang, D., 2021. Image recognition of wind turbine blade damage based on a deep learning model with transfer learning and an ensemble learning classifier. *Renew. Energy* 163, 386–397.
- Yu, J., He, Y., Liu, H., Zhang, F., Li, J., Sun, G., Zhang, X., Yang, R., Wang, P., Wang, H., 2023. An improved U-net model for infrared image segmentation of wind turbine blade. *IEEE Sens. J.* 23 (2), 1318–1327.
- Zampokas, G., Skartados, E., Alexiou, D., Tsiakas, K., Tzanakis, I., Roussos, N., Giakoumis, D., Kostavelis, I., Bouganis, C.-S., Tzovaras, D., 2022. WTA/TLA: a UAV-captured dataset for semantic segmentation of energy infrastructure. In: *2022 International Conference on Unmanned Aircraft Systems (ICUAS)*. pp. 552–561.