Image and Vision Computing 143 (2024) 104922

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Shadow detection using a cross-attentional dual-decoder network with self-supervised image reconstruction features

Ruben Fernandez-Beltran [a,*], Angélica Guzmán-Ponce [b], Rafael Fernandez [b], Jian Kang [c], Ginés García-Mateos [a]

[a] *Department of Computer Science and Systems, University of Murcia, 30100 Murcia, Spain*
[b] *Institute of New Imaging Technologies, University Jaume I, 12071 Castellón de la Plana, Spain*
[c] *School of Electronic and Information Engineering, Soochow University, Suzhou 215006, China*

ARTICLE INFO

Keywords:
Shadow detection
Semantic segmentation
Convolutional neural networks
Cross-attention
Dual-decoder

ABSTRACT

Shadow detection is a challenging problem in computer vision due to the high variability in lighting conditions, object shapes, and scene layouts. Despite the positive results achieved by some existing technologies, the problem becomes particularly challenging with complex and heterogeneous images where shadow-casting objects coexist and shadows can have different depths, scales, and morphologies. As a result, more advanced and accurate solutions are still needed to deal with this type of complexities. To address these challenges, this paper proposes a novel deep learning model, called the Cross-Attentional Dual Decoder Network (CADDN), to improve shadow detection by using fine-grained image reconstruction features. Unlike other existing methods, the CADDN uses an innovative encoder-decoder architecture with two decoder segments that work together to reconstruct the input images and their corresponding shadow masks. In this way, the features used to reconstruct the original input image can be used to support the shadow detection process itself. The proposed model also incorporates a cross-attention mechanism to weight the most relevant features for detecting shadows and skip connections with noise to improve the quality of the transferred features. The experimental results, including several benchmark image datasets and state-of-the-art detection methods, demonstrate the suitability of the presented approach for detecting shadows in computer vision applications.

1. Introduction

Shadow detection is a crucial component in many computer vision applications, as shadows cast on areas can greatly influence the accuracy of image analysis and machine vision algorithms. Typically, shadows form when an object obstructs light rays from a source, a common occurrence in both natural and man-made environments. Although shadows can provide valuable information on the relationship between objects and light sources [1], they often introduce challenges: creating false edges and boundaries, diminishing contrast and color information, and altering how objects appear. These effects can lead to difficulties in accurately identifying and classifying objects and scenes for algorithms [2]. Therefore, the role of single-image shadow detection is increasingly important. It focuses on automatically distinguishing shadowed areas from actual objects within images, aiming to improve the precision of tasks such as image segmentation, object detection, and recognition, among others [3–5].

The evolution of single-image shadow detection algorithms spans from traditional techniques based on handcrafted features to the latest advances in deep learning-based methods, all of which have been thoroughly researched and developed in the field [6–8]. Traditional approaches to shadow detection [9,10] typically utilize simple features that capitalize on shadow properties such as color, intensity, and texture. These methods are known for their computational efficiency and effectiveness in controlled settings. However, they often fail to handle variations common in real-world scenarios, such as changes in camera perspectives, time of day, or weather conditions. In contrast, deep learning-based techniques [11–13] demonstrate superior adaptability to complex and varied contexts. They achieve this through an end-to-end learning process, which makes them more robust and versatile compared to traditional methods. Specifically, deep learning models employ sophisticated neural network architectures, enabling them to extract and analyze image features at multiple levels of detail, thus capturing the more nuanced visual characteristics of shadows.

[*] Corresponding author.
E-mail address: rufernan@um.es (R. Fernandez-Beltran).

https://doi.org/10.1016/j.imavis.2024.104922
Received 12 May 2023; Received in revised form 18 January 2024; Accepted 26 January 2024
Available online 2 February 2024

In recent years, deep learning has significantly advanced single-image shadow detection by utilizing a variety of deep features. For example, some methods [12,13] employ multi-scale features to enrich contextual understanding of scenes. Others [11,14] integrate attention mechanisms to focus on important image areas, thus improving shadow detection. In addition, various models improve the robustness of detection by incorporating spatial or spectral signals, such as directional information [15] or intensity-based indicators [16]. Despite the adaptability and breadth of deep learning in shadow detection, most current models focus primarily on directly mapping input images to their shadow masks. This approach, although effective in many cases, can miss out on the advantages of using complementary image features to improve detection accuracy and robustness, especially in complex and varied data scenarios [12,17]. The widespread use of standard encoder-decoder frameworks may result in biased or incomplete deep features, as the decoder typically concentrates only on shadow mask reconstruction without considering other image features that could provide insight into complex interactions within the image [18]. In intricate situations, where shadows overlay across various planes and objects of differing scales and morphologies, this can lead to a complex mix of shadow characteristics, necessitating a more nuanced approach to accurately distinguish and understand these diverse shadow features.

Fig. 1 showcases two distinct scenarios in which shadows interact with various materials and structures, demonstrating the complexity of shadow morphologies. The first scenario, depicted in Fig. 1(a), highlights the interaction between the diffuse shadows of trees and the hard shadows of buildings. These shadows, which occur at different elevations, merge to form a composite shadow with varied shapes and sizes. This mix of shadow types presents a segmentation challenge, necessitating a deep understanding of the scene's structure to accurately differentiate between them. The second scenario, shown in Fig. 1(b), illustrates how shadows from different materials, such as a car bonnet and a tree, combine to create a diverse range of shadow appearances. This variety of shadow characteristics, or spectral heterogeneity, adds another layer of complexity to shadow detection. It requires a more refined approach to capture the intricate interactions of light with different structures and materials. Taking into account these complexities, it becomes evident that more sophisticated and robust deep learning-based methods are needed to effectively process and interpret these diverse and heterogeneous visual data in shadow detection.

To address these challenges, this research introduces the Cross-Attentional Dual-Decoder Network (CADDN), a novel model for single-image shadow detection. CADDN is specifically engineered to utilize detailed image reconstruction features, enhancing its ability to process complex data. What sets our model apart is its unique encoder-decoder architecture, which features two decoders: one for reconstructing the input image and another for generating the shadow mask. This design allows the shadow mask decoder to access and utilize the intricate details identified during image reconstruction. To ensure optimal feature transfer, CADDN incorporates a cross-attention mechanism. This mechanism selectively identifies and transfers only those features that are most relevant to shadow detection. Additionally, we improve the quality of the feature through noise-injected skip connections between the encoder and the image decoder. To effectively train CADDN, we have developed a new joint loss formulation. This formulation not only assesses the accuracy of the reconstructed images but also evaluates the precision of the shadow masks produced. In general, the key contributions of our work include:

1. We propose a novel deep learning model for shadow detection (CADDN) which is able to exploit fine-grained image reconstruction features to enhance shadow detection.
2. We define a joint loss formulation to train the proposed dual-decoder architecture.
3. We conduct a comparative analysis with different state-of-the-art detection methods and benchmark collections.
4. We empirically demonstrate the suitability of the proposed model for shadow detection.

Our shadow detection method goes beyond technical advances in computer vision to address practical challenges, aligning with the Sustainable Development Goals (SDGs) [19]. It contributes to SDG 9 (Industry, Innovation, and Infrastructure) by enhancing decision-making and resource use in urban planning, agriculture, and environmental monitoring through precise shadow detection. For SDG 11 (Sustainable Cities and Communities), it helps manage shadow-related issues in urban areas, like optimizing solar energy and cooling needs. In terms of SDG 13 (Climate Action), our method provides detailed shadow data for climate studies and understanding shadow impacts on local climates. Additionally, for SDG 15 (Life on Land), it improves remote sensing for tracking land changes, vegetation health, and wildlife habitats.

The remaining parts of this paper are organized as follows: Section 2 reviews related works and discusses their primary limitations when working with complex data. Section 3 presents the proposed shadow detection model, describing its network design and loss formulation. Sections 4 and 5 contain the experimental results and discussions. Finally, Section 6 concludes the paper and outlines some future research directions.

2. Related work

Since our work focuses on single-image shadow detection, this section is primarily dedicated to reviewing relevant studies in the single-image field. On the basis of their nature, single-image shadow detection techniques can be broadly classified into the following categories.

2.1. Traditional methods

Traditional approaches mainly involve hand-crafted features and localization methods for detecting shadows from images [7]. In this

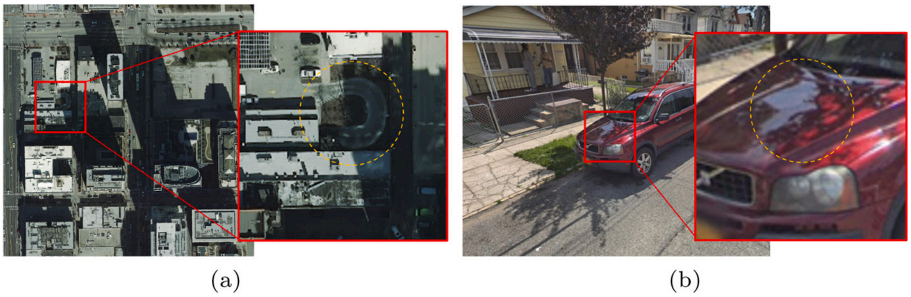


Fig. 1. Visual representation of shadow complexities. The figure on the left illustrates the interplay between diffuse tree shadows and hard shadows cast by buildings at varying heights, resulting in a composite of shadows with varied morphologies and scales. On the right, different materials, such as a car bonnet and a tree, contribute to a challenging spectrum of shadow types, showcasing the wide-ranging nature of shadow appearances and complexities.

way, it is possible to find in the literature different works investigating geometrical properties [20,21], color [9], edges [2], or even textures [10] to recognize shadows from static images. Despite their effectiveness in certain cases, most traditional methods tend to rely on explicit and distinguishable visual features that make them suffer from several limitations in natural scenes, for example, with soft or blurry shadows.

2.2. Deep learning methods

In contrast to traditional techniques, deep learning-based methods have recently shown significant improvements in shadow detection using large amounts of training data and powerful end-to-end convolutional neural networks (CNNs) [22]. One of the first studies in this area was conducted by Khan et al. [23], who developed a 3-layer CNN to learn the most relevant characteristics for shadow detection. Extending this idea, Vicente et al. [24] took advantage of noisy annotations to further improve the training process. Hosseinzadeh et al. [25] also incorporated a shadow prior, based on a support vector machine (SVM) classification of super-pixels, into the network input.

After all these seminal works, numerous other techniques have emerged in the literature. In general, the rationale behind deep learning-based shadow detection methods consists in learning a mapping function that transforms input images into their corresponding shadow masks. Hence, it is possible to identify some relevant deep learning technologies that can be instrumental in categorizing shadow detection networks:

- Encoder-Decoder Networks: These types of architecture (e.g. [26–28]) function through a two-step process: (i) the encoder, which compresses the input data into a condensed latent representation, and (ii) the decoder, which reconstructs the data from this compact form into the desired target. Using this network topology, deep learning-based shadow detection methods can learn complex transformations by encoding and then decoding the most relevant information through a trainable end-to-end process.
- Attention Networks: These techniques (e.g., [29–32]) dynamically adjust their focus, allocating increased attention to the salient features or image regions that are most relevant to the objective task. By weighting different parts of the input differently, attention mechanisms enhance the ability of shadow detection methods to recognize shadow patterns among other image components.
- Feature Fusion Networks: These networks (e.g., [33–36]) are designed to integrate and optimize information from different layers or inputs to capture a richer representation of the data. Therefore, shadow detection can be improved by exploiting a wide range of low-level details and high-level semantic features.
- Contextual Information Networks: These methods (e.g., [37–40]) integrate contextual information using global or local context aggregation modules. In the context of single-image shadow detection, this information certainly helps to better understand the relationships between shadows and their surroundings.
- Multi-Task Networks: These models (e.g., [41–46]) adopt a learning paradigm in which a single network is trained to perform multiple synergistic tasks simultaneously, such as shadow detection and removal or the interpretation of additional spatial cues. This approach leverages shared representations that can improve the efficiency and performance of each individual task. In shadow detection, some Generative Adversarial Networks (GANs) are clear examples, since they work to identify and remove shadows through a competitive process between the generator and discriminator networks.

Despite this broad categorization, it is important to note that almost all deep learning-based shadow detection methods, particularly the most recent ones, exhibit characteristics of several groups. This convergence of techniques within single frameworks demonstrates the trend of the field towards versatile models capable of nuanced discrimination and interpretation of shadows, as well as their relationship with the surrounding environment. Since the delineation of shadows from illuminated areas inherently divides images into semantically distinct segments, shadow detection aligns closely with semantic segmentation [47]. Therefore, the subsequent sections explore in detail relevant segmentation models and specialized deep learning models designed exclusively for single-image shadow detection.

2.2.1. General semantic segmentation networks

Being a very popular model, U-Net [48] was one of the most used encoder-decoder architectures for semantic segmentation, including shadow detection tasks [49]. In more detail, this network consists of a contracting path that captures context information and a symmetric expanding path that refines the localization of objects and regions from the input data. Due to the high effectiveness and versatility of this architecture, different extensions can be found in the literature. In [50], the authors proposed U-Net++, which employs a nested design with multiple skip pathways at different resolutions to provide better segmentation results. Liu et al. [51] also developed the pyramid attention network (PAN) which extends the U-Net with a novel pyramid attention module. More specifically, the defined attention module has a pyramidal structure that covers multiple spatial resolution scales. In addition, standard skip connections are modified with such attention blocks to reduce the semantic gap between the encoder and decoder paths. Similarly, Fan et al. [52] presented MaNet, which introduced different attention modules to focus on the most important features at each scale level, with the objective of enhancing the accuracy and robustness of the model. Chaurasia et al. [53] also developed LinkNet which mainly replaced standard U-Net skip connections by residual ones in order to improve gradient flow and make training more stable.

Despite the success of these variants, other authors decided to explore alternative designs. For example, it is the case of Seferbekov et al. [54] who created the feature pyramid network (FPN) to exploit features from different scales. Specifically, the FPN consists of a bottom-up pathway that generates features at different resolutions, and a top-down pathway that works at the same resolution as the input image and combines features from different scales using lateral connections. Zhao et al. [55] also formulated the pyramid scene parsing network (PSPNet) which uses a pyramid pooling module that divides feature maps into different regions and applies pooling operations at multiple scales, allowing the network to capture global and local contextual information. Similarly, other studies [56,57] suggested the use of dilated convolutions to increase receptive fields while effectively capturing multi-scale contextual information. More recently, some other works have taken advantage of the so-called visual transformer networks [58]. It is the case of Xie et al. [59] who defined a transformer network for semantic segmentation (SegFormer) by exploiting long-range dependencies and spatial relationships with the so-called multi-head self-attention mechanism. The authors in [60] also proposed a hierarchical extension (HiFormer) that hierarchically aggregates features across multiple levels of the encoder with an FPN-based architecture to further refine feature maps.

2.2.2. Specialized shadow detection networks

In addition to the methods mentioned above, the literature also includes specialized deep learning models exclusively designed for single-image shadow detection. These approaches focus on tailoring architectures and learning mechanisms specifically for shadow detection, with the aim of achieving improved accuracy and robustness in the process. For example, it is the case of Zhu et al. [11] who presented the bidirectional feature pyramid network with recurrent attention residual modules (BDRAR) for single-image shadow detection. In particular, this specialized network leverages two key elements: an attention module to refine context features from adjacent CNN layers, and a bidirectional feature pyramid network to aggregate shadow contexts from different

CNN layers. By refining the context features in both directions, the BDRAR method demonstrated the ability to reduce false predictions while improving shadow details. Following a similar motivation, Hu et al. [15] developed the direction-aware spatial context network (DSC), which learns attention weights when aggregating spatial context features to recover direction-aware contexts for detecting shadows. In [13], Fang et al. proposed a shadow detection network based on effective shadow contexts (ECA). Specifically, the ECA method was designed to capture global and local contexts of shadows using a multi-scale encoder-decoder structure with attention modules. In addition, it employs a shadow refinement module to enhance the boundaries of the predicted shadows.

Similarly, other researchers have explored alternative ways of exploiting spatial contexts and attention mechanisms in their methods. For example, Liu et al. [14] designed the multi-scale spatial attention network (MSASDNet) which focuses on extracting features at different spatial scales while applying spatial attention to each scale in order to reduce the interference of non-shadow regions at each spatial level. In [12], the authors presented the fast shadow detection network (FSDNet) based on three different modules: inverted residual bottlenecks to extract multi-scale features, a direction-aware spatial context module to provide global context information, and a detail enhancement module to refine low-level featured details based on the distance between low-level and high-level feature maps. Another relevant approach can be found in [61], where Xie et al. proposed the omni-scale global–local aware network (OglaNet). In more detail, the OglaNet method consists of two main components: a global–local network to extract global and local convolutional features, and a pyramid pooling module for capturing multi-scale contextual information with different pooling sizes. In this way, the outputs of both modules are concatenated and fed into a final convolutional layer to generate the final shadow predictions.

Certainly, spatial contexts and attention mechanisms have proven to be important elements in shadow detection networks. However, other researchers have considered and explored additional shadow cues in their approaches. In the case of [16], Zhu et al. introduced a feature decomposition and reweighting network (FDRN) specifically designed to address the intensity bias problem in shadow detection. Since deep features are generally sensitive to intensity values, the authors proposed decomposing the uncovered features into intensity-variant and intensity-invariant components and reweighting each part to control such intensity effect. Following a similar motivation, Zhu et al. [62] developed a complementary mechanism for jointly exploiting the information extracted from shadowed and non-shadowed regions. Specifically, the proposed method makes use of two interactive branches: one for predicting shadow masks, and another for their complementary masks (representing non-shadows). In this way, the deactivated intermediate features of one branch can be delivered to the other through a negative activation technique. Other researchers have also explored different types of complementary information. In [63], Wu et al. took advantage of the inherent noise of shadow image collections to develop a robust detection scheme, which identifies the most reliable samples and propagates this information using graph convolutional networks. Jie et al. [64] also demonstrated the benefits of incorporating a randomized feature sampling strategy into their transformer-based shadow detection network.

Building on these explorations, further advancements in the field continue to emerge. Jie et al. assessed the Segment-Anything Model (SAM) for shadow detection, revealing its limitations compared to specialized models [26]. Jiao et al. proposed the Refined UNet-v4 for edge-refined cloud and shadow detection in remote sensing images [27]. A paper by [28] introduced an encoder-decoder network with a channel-attention module for remote sensing images, focusing on shadow characteristics. Kumar et al. developed SEAT-YOLO, a YOLO-based architecture with squeeze-excite and spatial attention modules for shadow detection [29]. Zhou et al. presented an improved method to detect face shadows using channel and spatial attention [30]. Liu et al. introduced SCOTCH and SODA, a transformer-based framework for video shadow detection, addressing shadow deformations and contrastive learning [31]. Yucel et al. proposed LRA and LDRA for efficient shadow detection and removal, focusing on shadow region reconstruction [32].

Moving forward with other very recent advancements, the shadow detection field continues to evolve. Cong et al. developed a shadow detection network with a style-guided dual-layer disentanglement architecture [33]. Zhang et al. introduced CIFNet, a multi-supervised feature fusion attention network for cloud and shadow detection [34]. Another study by Zhang et al. proposed CRSNet, employing multiple modules for cloud and shadow detection in remote sensing imagery [35]. Feng et al. designed OAMSFNet, an orientation-aware network using pseudo-shadow information for remote sensing images [36]. Chen et al. presented a boundary-aware network for enhanced shadow detection [37]. Wu et al. combined uncertainty analysis with a GCN-based strategy for single-image shadow detection [38]. Zhang et al. proposed a method combining neighborhood similarity and intensity information for shadow detection in video SAR [40]. Valanarasu et al. introduced a method leveraging shadow removal for fine-context shadow detection [41]. Zhang et al. explored residual and illumination with GANs for shadow detection and removal [42]. Another study by Zhang et al. proposed SpA-Former, a transformer-based network for shadow removal [43]. Guo et al. introduced ShadowDiffusion, a diffusion framework for shadow removal [44]. Ahn et al. combined a shadow transformer network with GANs for domain adaptation in shadow removal [45]. Lastly, Xu et al. proposed a dynamic convolution module for shadow detection and removal [46].

All of these studies represent a diverse range of approaches and innovations in the field of shadow detection, highlighting the ongoing advances and varied methodologies employed to tackle this complex challenge.

2.3. Novelty of the work

Undoubtedly, deep learning models have revolutionized shadow detection, enabling efficient and effective solutions. However, accurately segmenting shadows from highly heterogeneous and complex data still poses significant challenges to the scientific community [12,17]. While some methods exploit low intensity as a strong indicator of shadows [16], others try to enhance the process by using additional shadow cues, such as shadow/non-shadow correspondences or region connectivity [14,61,62]. However, existing architectures are generally unable to prioritize fine-grained image reconstruction features that, without being directly related to shadow masks, may help to improve the detection process. Deep learning-based shadow detection models commonly use a standard encoder-decoder network to predict shadow masks from input images, e.g. [11,12]. However, this approach may neglect some fine-detailed image features that are not directly related to shadows but could still influence their accurate estimation by providing a better understanding of the scene. In other words, the standard decoder path for predicting shadow masks may fail to consider some image features that could help to understand complex interactions among light, structures, and materials, helping to detect some complex cases. Note that complex scenarios require different abstraction levels to handle different planes, scales, and morphologies of objects that generate different shadow characteristics, which can be mixed in the scene and need to be identified with fine-grained image features.

In response to these challenges, this paper presents a novel encoder-decoder network for shadow detection that takes advantage of image reconstruction features using a dual image-shadow decoder topology. Our methodology deviates from the standard singular focus on shadow masks by incorporating a secondary pathway dedicated to reconstructing the non-shadow portions of the image. In addition, it also deviates from other existing dual-path or multitask networks (e.g. [41,62]) by selectively transferring and exploiting image reconstruction features without the need of using any other additional data, such as in the case of joint shadow detection/removal methods that also require clean images. The use of skips connections with noise and cross-attention enable our network to improve the quality of the transferred features to better grasp the nuances of both shadowed and illuminated regions, positioning our work as an advancement over other existing approaches. Experimentally, our model demonstrates improved robustness in scenes with intricate interplays of light and texture. Furthermore, ablation studies underline the effectiveness of each architectural component, particularly the reconstruction features, which prove pivotal in discerning subtle shadow details.

3. Methodology

This section presents the proposed shadow detection model. First, let us introduce the considered notation. Let $I_{\mathrm{IN}} \in \mathbb{R}^{(H\times W\times B)}$ represent an input image with B spectral bands and a $(H \times W)$ spatial size. The corresponding ground-truth shadow mask is identified by $I_{\mathrm{GT}} \in \mathbb{B}^{(H\times W\times 1)}$, while $I_{\mathrm{SH}} \in \mathbb{B}^{(H\times W\times 1)}$ denotes the shadow mask predicted by the network. Finally, the reconstructed output image is represented by $I_{\mathrm{OUT}} \in \mathbb{R}^{(H\times W\times B)}$. Under this framework, the proposed network aims to approximate a mapping function $\mathscr{F}$ (Eq. (1)), using a supervised learning approach that relies on ground-truth data for loss computations.

$$\mathscr{F}(I_{\mathrm{IN}}) = (I_{\mathrm{SH}}, I_{\mathrm{OUT}}) \tag{1}$$

3.1. CADDN: Cross-attentional dual-decoder network for shadow detection

Fig. 2 illustrates our newly developed CADDN model. This architecture stands out with a dual-decoder design, integrating an encoder backbone ($\mathscr{E}$) with two specialized decoder segments: the Shadow-Decoder ($\mathscr{D}_S$) and the Image-Decoder ($\mathscr{D}_I$). These segments work in tandem, with $\mathscr{D}_S$ dedicated to estimating shadow masks and $\mathscr{D}_I$ focused on reconstructing the input data. The core concept of this configuration is to harness the power of image reconstruction features to enhance the shadow detection process.

The process begins with the encoder backbone $\mathscr{E}$, which compresses the input images into a condensed feature space. This embedded representation then feeds into the two decoders. The Image-Decoder ($\mathscr{D}_I$) takes on the task of reconstructing the input images from this feature space, while the Shadow-Decoder ($\mathscr{D}_S$) focuses on generating the corresponding shadow masks. A key feature of our model is its ability to transfer crucial features from $\mathscr{D}_I$ to $\mathscr{D}_S$. This transfer is instrumental in enhancing shadow detection capabilities.

The designed dual-decoder approach goes beyond mere parallel processing, focusing instead on creating a synergistic interaction between the decoders. By training these decoders simultaneously in an end-to-end manner, CADDN effectively leverages detailed image reconstruction features to improve shadow detection. This aspect is particularly beneficial in complex scenarios where shadows display varied morphologies, scales, and spectral characteristics. The subsequent sections will dive deeper into each component of CADDN, specifically focusing on the Encoder ($\mathscr{E}$), Shadow-Decoder ($\mathscr{D}_S$), and Image-Decoder ($\mathscr{D}_I$) as marked in Fig. 2.

3.1.1. Encoder network ($\mathscr{E}$)

The encoder backbone, denoted $\mathscr{E}$, aims to transform the input image into a more compact and representative characterization that can preserve the most meaningful features from a shadow detection perspective. Specifically, we built $\mathscr{E}$ on top of the popular ResNet-34 [65] architecture due to its excellent balance between model complexity and computational efficiency in shadow detection and other analogous segmentation applications [39,66]. A graphical representation of the encoder network is provided in Fig. 3, where the following abbreviations are used: Conv2D (convolutional layer), BN (batch norm), ReLU (rectified lineal unit), MaxPool (max pooling), Add (residual layer), E_H (head features), E_{S1} (stage-1 features), E_{S2} (stage-2 features), E_{S3} (stage-3 features), and E_{S4} (encoding space features). Like in ResNet-34, the defined encoder begins with a convolutional layer containing 64 filters with a kernel size of 7 × 7, followed by a max-pooling that halves the spatial dimensions. After this, there are four stages of residual blocks, each one consisting of two convolutional layers with a kernel size of 3 × 3 and a residual connection. The first stage contains three residual

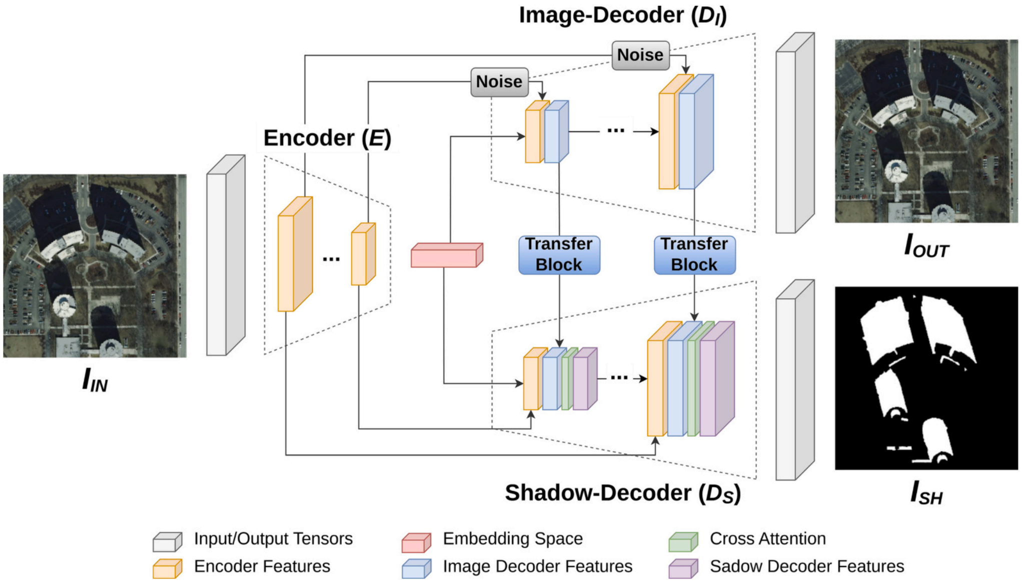


Fig. 2. Overview of the proposed Cross-Attentional Dual-Decoder Network (CADDN). The figure illustrates how the input images (I_{IN}) are encoded by means of $\mathscr{E}$ into an embedding space. Then, two connected decoder segments (image-decoder $\mathscr{D}_I$ and shadow-decoder $\mathscr{D}_S$) are used to produce a reconstructed version of the inputs (I_{OUT}) and estimate the corresponding shadow masks (I_{SH}). Note that skip connections with noise are integrated between $\mathscr{E}$ and $\mathscr{D}_I$ to improve the robustness of the decoded image features. Additionally, transfer blocks are used between $\mathscr{D}_I$ and $\mathscr{D}_S$ to adapt the multi-modal nature of both decoder streams. Finally, $\mathscr{D}_S$ incorporates a cross-attention module to allow attending to different feature sequences regardless of whether they come from the encoder, image-decoder or shadow-decoder paths.

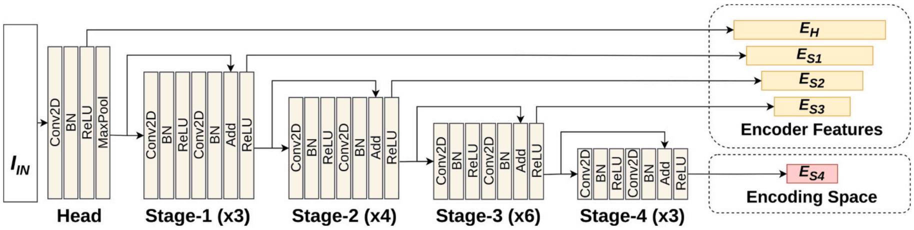


Fig. 3. Encoder network ($\mathscr{E}$). This diagram illustrates the encoder component of the proposed architecture. Based on the ResNet-34 [65], the defined encoder employs successive stages of convolutional layers (Conv2D), batch normalization (BN), rectified linear units (ReLU), max pooling (MaxPool) and residual connections (Add) to progressively build a hierarchy of features from the head (E_H) to the deepest encoding space (E_{S4}). This progression through stages utilizes increasing filter sizes to refine feature maps, facilitating the feature transfer to subsequent decoder segments dedicated to shadow mask prediction and image reconstruction.

blocks with 64 convolutional filters each. In the second stage, there are four residual blocks, each with 128 filters. The third stage comprises six residual blocks with 256 filters each. Finally, the last stage consists of three residual blocks with 512 filters, each one of them. Note that strided convolutions are used in the transition between stages to reduce the spatial dimensions of the feature maps. In this way, the encoder is able to produce a set of feature maps, as shown in Eq. (2), with reduced spatial dimensions and progressively complex representations that can be fed into the two decoder segments.

$$\mathscr{E}(I_{\mathrm{IN}}) = (E_H, E_{S1}, E_{S2}, E_{S3}, E_{S4}) \tag{2}$$

3.1.2. Image-decoder network ($\mathscr{D}_I$)

The proposed image-decoder segment ($\mathscr{D}_I$) seeks to reconstruct the original input image from the features generated by the encoder backbone. Therefore, the projection learned by $\mathscr{D}_I$ should capture fine-grained visual details necessary to produce an accurate reconstruction of the input image from the compressed embedding space. To enable the use of multi-resolution information [14,61], we adopt a U-Net decoder shape that considers the set of previously encoded features. However, it is important to note that the use of standard U-Net skip connections may not be optimal, since simply bypassing initial high-resolution features from the encoder could compromise the relevance of lower-resolution features. To address this issue, we introduce a dropout noise rate of η to the encoder feature maps used in the shortcut connections. Fig. 4 shows the architecture of the proposed image-decoder, where the following abbreviations can be found: UP (up-sampling layer), Cat (concatenation layer), TanH (hyperbolic tangent function), D_H (head image-decoder features), D_{S1} (stage-1 image-decoder features), D_{S2} (stage-2 image decoder features), and D_{S3} (stage-3 image-decoder features). As can be seen, the expansive path consists of five up-sampling blocks (i.e. Stage-3i, Stage-2i, Stage-1i, Head-i, and Tail-i), each containing two 3 × 3 convolutions with a decreasing number of filters (i.e., 256, 128, 64, 32, and 16, respectively) to gradually rearrange the features from a channel-wise representation to a spatial representation. This design enables the generation of image-decoder features that are symmetric to the encoder ones, and thus, they can be jointly exploited by the shadow-decoder segment. Finally, an image reconstruction block with four more convolutions is used to re-project the obtained features onto the original image domain. The whole process of the defined image-decoder can be represented by Eq. (3).

$$\mathscr{D}_I(\mathscr{E}(I_{\mathrm{IN}}), \eta) = (D_{S3}, D_{S2}, D_{S1}, D_H, I_{\mathrm{OUT}}) \tag{3}$$

3.1.3. Shadow-decoder network ($\mathscr{D}_S$)

Regarding the shadow-decoder, $\mathscr{D}_S$ is a critical component of the proposed architecture, as its objective is to generate the final prediction of shadows by using the features extracted by the encoder and image-decoder networks. By incorporating the high-level semantics of the encoder features (i.e., E_H, E_{S1}, E_{S2} and E_{S3}) and the fine-grained visual details of the image-decoder features (i.e., D_H, D_{S1}, D_{S2} and D_{S3}), the shadow-decoder can be boosted to handle more complex and diverse conditions for shadow detection. To achieve this, we employed a U-Net shape with dual-skip connections, allowing access to both feature paths (encoder and image-decoder features) when identifying shadows. Furthermore, transfer and attention modules are incorporated to further enhance the robustness of the shadow predictions.

The architecture of the proposed shadow-decoder is shown in Fig. 5, where the same abbreviations as above are used. As can be observed, it consists of five up-sampling stages (i.e., Stage-3 s, Stage-2 s, Stage-1 s, Head-s, and Tail-s), each comprising two 3 × 3 convolutions and a progressively reduced number of filters (i.e., 256, 128, 64, 32, and 16,

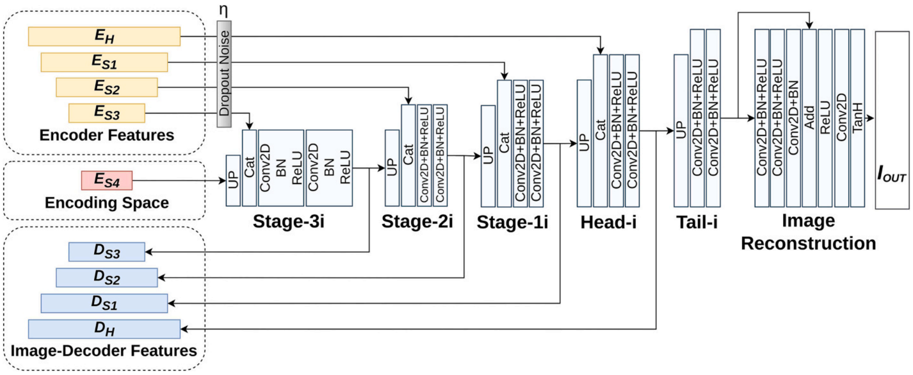


Fig. 4. Image-decoder network ($\mathscr{D}_I$). This figure delineates the image-decoder segment of our architecture, responsible for reconstructing the input image from the encoded features. It mirrors the U-Net decoder structure and advances from the deepest encoding space (E_{S4}) through up-sampling layers (UP) and concatenations (Cat) with the encoder features, which have been selectively perturbed with dropout noise to enhance feature quality. The process incorporates repeated sequences of convolution (Conv2D), batch normalization (BN), and rectified linear activation (ReLU), systematically restoring the resolution of feature maps. The final output passes through a hyperbolic tangent activation (TanH) to produce the final reconstructed image. The symmetry with the encoder allows a comprehensive multi-level feature exploitation, crucial for the accurate image reconstruction.

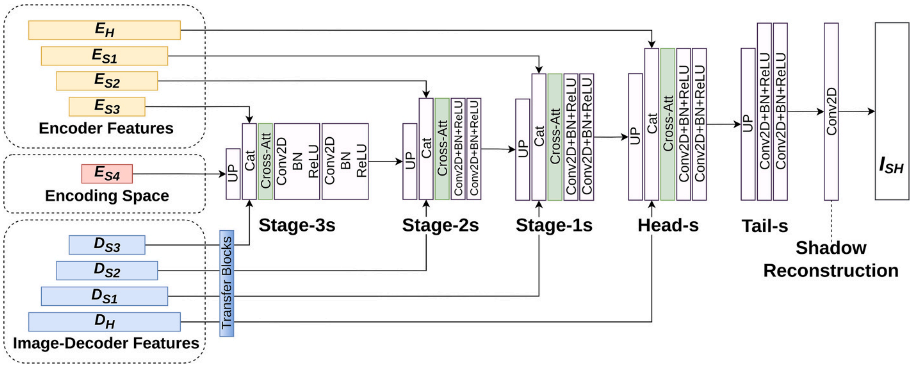
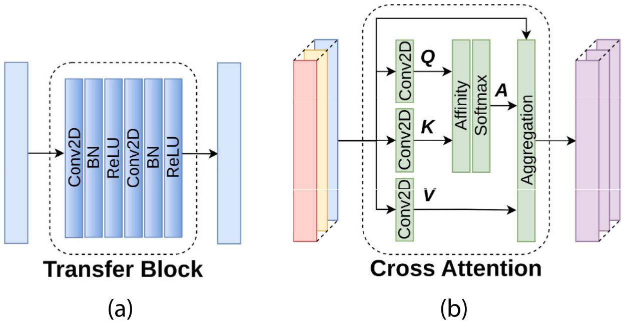


Fig. 5. Shadow-decoder network ($\mathscr{D}_S$). This scheme depicts the shadow-decoder component of our architecture, designed to predict shadow masks from features refined by both the encoder and the image-decoder. It follows a U-Net-inspired structure, leveraging up-sampling stages with transfer blocks and cross-attention to integrate multi-level encoder features (E_H, E_{S1}, E_{S2} and E_{S3}) with detailed image-decoder features (D_H, D_{S1}, D_{S2} and D_{S3}). This strategic assembly facilitates the ability to handle a wider variety of complex shadow conditions. The defined stages utilize up-sampling (UP) and concatenation layers (Cat) together with dual convolutional sequences (Conv2D) followed by batch normalization (BN) and ReLU activation to progressively shape shadow features.

respectively). Additionally, a final convolution, identified as the shadow reconstruction block, is included to generate the corresponding output shadow masks. Essentially, the architecture of the proposed shadow-decoder resembles that of the previously described image-decoder, but with three main differences. First, the features learned by the image-decoder are fed into the concatenation layer (Cat), after being processed by a transfer block. Second, a cross-attention module is incorporated to allow the shadow-decoder to attend to different feature sequences regardless of whether they come from the encoder, image, or shadow paths. Third, the final reconstruction block is simplified to a single convolution due to the simplicity of the output, which is a shadow mask. Eq. (4) formulates the shadow-decoder process and Fig. 6 provides a visual description of the transfer blocks and cross-attention modules considered.

$$\mathscr{D}_S(\mathscr{E}(I_{\text{IN}}), \mathscr{D}_I(\mathscr{E}(I_{\text{IN}}), \eta)) = (I_{\text{SH}}) \tag{4}$$

In more detail, each transfer block (Fig. 6(a)) is made up of two 3 × 3 convolutions with a number of filters equal to the input channels. This configuration is used to allow for better adaptation of the image-decoder features that come from a multi-modal stream. The considered cross-attention module (Fig. 6(b)) is based on the so-called multi-head attention mechanism of transformer networks, which has been successfully applied in several semantic segmentation applications [59,60,67]. Assuming a generic input feature map $M \in \mathbb{R}^{(s_1 \times s_2 \times s_3)}$, the cross-attention module begins with two 1 × 1 convolutions to create a lower-dimensional embedding for queries (Q) and keys (K), where $(Q, K) \in \mathbb{R}^{(s_1 \times s_2 \times \lfloor s_3/8 \rfloor)}$. Next, the affinity matrix between Q and K is computed by iterating the spatial dimensions as follows. At each spatial position p, query features are extracted in depth as $Q_p \in \mathbb{R}^{(\lfloor s_3/8 \rfloor)}$. Similarly, all the feature vectors in K that are in the same row or column with p are extracted as $K_{i,p} \in \mathbb{R}^{(s_1+s_2-1 \times \lfloor s_3/8 \rfloor)}$. Then, the corresponding affinity degrees are computed as:



(a) (b)

Fig. 6. Transfer and Cross-Attention modules in CADDN. The image on the left shows the proposed transfer block, which employs a sequence of convolution (Conv2D), batch normalization (BN), and rectified linear activation (ReLU) layers to refine the image-decoder features, ensuring their compatibility for precise shadow detection. The second image depicts the cross-attention mechanism, a module that utilizes transformer network principles with queries (Q), keys (K) and values (V), to selectively concentrate on features with high affinity across multiple data streams.

$$d_{i,p} = Q_p K_{i,p}^T \tag{5}$$

where $d_{i,p} \in D$ is the correlation between the query features in the spatial position p and the key features in the same row/column as p, such that $D \in \mathbb{R}^{(s_1+s_2-1 \times s_1 \times s_2)}$. After this operation, a soft-max layer is used to transform D into a normalized attention map A, which will be aggregated to the input features. To achieve this, the input (M) is processed by a 1 × 1 convolution to generate the corresponding value embedding $V \in \mathbb{R}^{(s_1 \times s_2 \times s_3)}$ for feature adaptation. Then, at each spatial position p, feature vectors belonging to the same row/column as p are extracted in $V_{i,p} \in \mathbb{R}^{(s_1+s_2-1 \times s_3)}$. With this, the aggregation step can be formulated as Eq. (6) shows, being $M_{p'}$ the feature vector at position p on the output feature map $M' \in \mathbb{R}^{(s_1 \times s_2 \times s_3)}$.

$$M_{p'} = \sum_{i=1}^{s_1+s_2-1} A_{i,p} V_{i,p} + M_p \tag{6}$$

3.2. *Proposed joint loss formulation*

In this section, we present the loss formulation used for training the proposed model. It should be noted that, as a supervised method, paired data volumes $(I_{\text{IN}}, I_{\text{GT}})$ are required for the training process. As illustrated in Fig. 2, the proposed approach adopts a dual-decoder architecture that requires joint optimization. Therefore, we consider the following loss terms for training: image reconstruction ($\mathscr{L}_I$) and shadow matching ($\mathscr{L}_S$). We will now provide a detailed description of them:

($\mathscr{L}_I$) The first loss is dedicated to optimizing the image-decoder segment ($\mathscr{D}_I$) to ensure a high-quality reconstruction of the input image (I_{OUT}). When it comes to full-reference image quality metrics, the structural similarity index (SSIM) [68] is one of the most popular choices for image reconstruction because it provides a measure of structural similarity between two images based on patches, rather than simply measuring the difference in the values of the pixels [69]. Eq. (7) shows its mathematical formulation, where x and y are overlapped image patches, $\mu_{(\cdot)}$ represents their corresponding means, $\sigma_{(\cdot)}$ the standard deviations, $\sigma_{(\cdot,\cdot)}$ the cross-covariance, and $c_{(\cdot)}$ constants for numerical stability. Since the averaged result incorporates both local and global information, the SSIM metric becomes particularly robust for managing variations in illumination, contrast, and other features that certainly play a key role in shadow detection. In addition, its bounded nature [70] also makes this index a suitable election for a multi-loss optimization scenario. Therefore, we adopt the SSIM as the figure of merit of our image reconstruction loss. Eq. (8) details the considered expression, where Ω represents the window grid, |Ω| is the total number of patches in the image domain, and $I^p_{(\cdot)}$ the patches extracted from the corresponding images.

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{\left(\mu_x^2 + \mu_y^2 + c_1\right)\left(\sigma_x^2 + \sigma_y^2 + c_2\right)} \tag{7}$$

$$\mathscr{L}_I(I_{\mathrm{OUT}}, I_{\mathrm{IN}}) = \frac{1 - \left(\frac{1}{|\Omega|}\sum_{p\in\Omega}\mathrm{SSIM}(I^p_{\mathrm{OUT}}, I^p_{\mathrm{IN}})\right)}{2} \tag{8}$$

($\mathscr{L}_S$) The objective of the second loss consists in fitting the predictions made by the shadow-decoder (I_{SH}) to their corresponding ground-truth shadow masks (I_{GT}). In the context of deep learning-based semantic segmentation [71], the Dice loss [72] is a widely used function because it deals with the class imbalance problem, which is a common issue in shadow detection [73]. More specifically, this loss is based on the Dice coefficient, which measures the similarity between two sets by computing the ratio of their intersection to their union. Eq. (9) shows its mathematical formulation, where X and Y are two binary masks, $|\cdot|$ denotes the cardinality operator, and c is a small constant for numerical stability. As is possible to see, this coefficient is a bounded score, with values ranging from 0 (no overlap) to 1 (perfect overlap), that penalizes false negatives and false positives equally. Accordingly, we use the Dice loss presented in Eq. (10) for training our shadow-decoder.

$$\mathrm{Dice}(X,Y) = \frac{2|X\cap Y| + c}{|X| + |Y| + c} \tag{9}$$

$$\mathscr{L}_S(I_{\mathrm{SH}}, I_{\mathrm{GT}}) = 1 - \mathrm{Dice}(I_{\mathrm{SH}}, I_{\mathrm{GT}}) \tag{10}$$

By combining these two terms, we define a joint loss formulation for the proposed CADDN model that works for simultaneously optimizing the reconstruction of the input image and the prediction of its shadow mask. To allow fine-tuning the balance between both aspects, we use a hyper-parameter β that controls the relative importance of the image reconstruction loss. Eq. (11) shows the expression of the proposed joint loss, which provides a unified framework for optimizing the proposed decoder segments and encourages the model to learn representations that capture both fine-grained image structure and shadow information.

$$\mathscr{L}_{\mathrm{CADDN}} = \mathscr{L}_S(I_{\mathrm{SH}}, I_{\mathrm{GT}}) + \beta\mathscr{L}_I(I_{\mathrm{OUT}}, I_{\mathrm{IN}}) \tag{11}$$

4. Experiments

This section presents the experimental part of the work, describing the datasets (Section 4.1), settings (Section 4.2), and results (Section 4.3). To offer a more comprehensive evaluation of the proposed method, additional experiments are presented in the appendices. These experiments include a parameter sensitivity analysis (Appendix A), an ablation study (Appendix B), and a trade-off analysis (Appendix C).

4.1. Datasets

The following image collections were considered in the experiments:

1. AISD [74]: The Aerial Imagery dataset for Shadow Detection (AISD) [74] consists of 514 images, extracted from the Inria Aerial Image Labeling Dataset [75]. The collection covers a wide range of complex scenarios, including dense metropolitan areas, residential neighborhoods, industrial regions, and rural resorts, making it a suitable dataset for evaluating shadow detection algorithms. The authors provide the data partitioned into three sets, training (412 images), validation (51 images), and test (51 images), along with their corresponding ground-truth shadow masks. All the images are RGB shots with a size of 512 × 512 and a spatial resolution of 0.3 meters per pixel.
2. CUHKMAP [12]: This dataset is part of the Chinese University of Hong Kong (CUHK) archive, which combines images from various shadow detection repositories. The CUHKMAP comprises a total of 1595 complex scenes collected from Google Street View, each manually labeled for shadow regions. The dataset includes panoramic views captured from different viewpoints and under various lighting conditions, making it particularly challenging for shadow detection. The data are divided into three partitions: training (1116 images), validation (160 images), and test (319 images). It is worth noting that the images in the dataset are RGB shots and have varying sizes, typically around 400 × 600 pixels.
3. SBU [24]: The Stony Brook University (SBU) is another reference shadow detection collection that includes 4723 images collected from the MS COCO dataset [76] and the Web. The included scenes are highly diverse with different types of environment and objects, providing a challenging benchmark for evaluating shadow detection algorithms. The dataset is organized into two parts, training (4085 images) and test (638 images), both with their manually annotated shadow masks. Since no validation is available in this case, we use a fixed partition with 10% of the training for validation purposes. The images in the SBU dataset are RGB shots with different resolutions and aspect ratios, having a sample size around 400 × 400.

4.2. Experimental settings

To validate the proposed shadow detection model, we perform an experimental comparison against several popular deep learning-based semantic segmentation and shadow detection methods available in the literature. In the category of general-purpose semantic segmentation, we consider U-Net++ [50], MaNet [52], LinkNet [53] and HiFormer [60]. As specialized shadow detection networks, we include FSDNet [12], MSASDNet [14], OglaNet [61], BDRAR [11], DSC [15], ECA [13] and FDRN [16].

In the experimental setup to evaluate the proposed CADDN model and the other shadow detection methods, we maintained uniformity across all models for a fair comparison. This uniformity included using the same datasets and settings as provided by the authors of each method. Specifically, for the proposed CADDN model, we opted for a default configuration with $\eta = 0.1$ and $\beta = 1$. To standardize the feature extraction and optimization process across different models, the Dice loss function and the ResNet-34 backbone were used whenever applicable. We intentionally avoided using data augmentation techniques, pre-trained weights, or additional post-processing steps. This decision was made to ensure that the comparison focused solely on the architectures' performance without external enhancements. To provide a robust statistical analysis, we performed five Monte Carlo runs for each model and reported the average results.

For training details, all datasets were resized to a uniform resolution of 224 × 224 pixels. This resizing served two purposes: (i) it made the input data compatible with all models, and (ii) it helped manage memory requirements. We used the ADAM optimizer for training, setting the training duration to 100 epochs with an initial learning rate of $1e^{-3}$. To facilitate learning, we implemented a learning rate decay strategy, reducing it by 50% every 20 epochs. The batch size was set to 4 for all models. During the training phase, we continuously monitored the model performance in the validation set. The model instance that performed the best in the validation set was saved for subsequent evaluation in the test set. This approach ensured that we used the most effective version of each model for the final testing and comparison.

In our evaluation process, we employed four distinct metrics to quantitatively assess the performance of the models: Intersection over Union (IoU), F1-score (F1), overall accuracy (OA), and Balanced Error Rate (BER). These metrics provide a comprehensive analysis of the models' effectiveness in shadow detection from different perspectives. Additionally, we complemented our quantitative assessment with qualitative analysis, examining several visual results to gain insight into the practical performance of the models. For conducting the experiments, our computational setup included a system equipped with an Intel(R) Core(TM) i5–11,400 processor, NVIDIA GeForce RTX 3060 graphics card, and 64 GB of DDR4 RAM. The system ran on Ubuntu 20.04 (64-bit version) and utilized Pytorch 1.6.0 with CUDA 10.1 for efficient processing and model training. To facilitate reproducibility and further research, the code developed for this study will be made available at https://github.com/rufernan/CADDN.

4.3. Results

Tables 1, 2 and 3 present the quantitative evaluation results achieved on the three benchmark shadow detection datasets. Each table shows the considered methods in rows and the selected evaluation metrics in columns, with the best result highlighted in bold. The optimal values for the reported metrics are IoU (1), F1 (1), OA (100%), and BER (0%). For visual evaluation purposes, Figs. 7 8 and 9 show some examples of the shadow detection results obtained on images from the three datasets.

4.4. Computational complexity

This section presents the results of three key performance metrics to analyze the computational cost of the considered shadow detection models. Specifically, we focus on the training time in seconds (Fig. 10), test time in seconds (Fig. 11), and the maximum demand for GPU memory in megabytes (Fig. 12). Since a uniform experimental setting has been consistently employed across all datasets, the results presented in this section are specifically for the AISD collection.

Based on these computational results, the proposed approach, CADDN, performs competitively compared to other state-of-the-art methods. The training time of CADDN is similar to that of some other models, such as MaNet and FSDNet, and is notably faster than several others, including HiFormer, MSASDNet, and ECA. Furthermore, the test time of CADDN is relatively efficient, falling within the same range as that of models like U-Net++, MaNet, and LinkNet. It is worth highlighting that, despite the complexity and capabilities of CADDN, it does not require excessive computational resources. In terms of GPU memory usage, CADDN utilizes a moderate amount of memory compared to some other models like OglaNet and MSASDNet, which demand more GPU resources. These facts indicate that CADDN strikes a good balance between computational efficiency and performance.

Table 1
Quantitative performance analysis on the AISD dataset. This table presents a comprehensive comparison of various shadow detection methods, including our proposed CADDN model, evaluated on the AISD dataset. Metrics include Intersection over Union (IoU), F1 Score, Overall Accuracy (OA), and Balanced Error Rate (BER).

Methods	IoU	F1	OA (%)	BER (%)
U-Net++ [50]	0.8374	0.9111	96.15	5.83
MaNet [52]	0.8387	0.9119	96.16	5.64
LinkNet [53]	0.8343	0.9092	96.05	5.76
HiFormer [60]	0.8010	0.8889	95.36	8.02
FSDNet [12]	0.7832	0.8777	94.79	8.15
MSASDNet [14]	0.8341	0.9091	96.04	5.79
OglaNet [61]	0.8363	0.9105	96.15	6.07
BDRAR [11]	0.7888	0.8813	94.82	7.33
DSC [15]	0.7978	0.8869	95.04	6.49
ECA [13]	0.8354	0.9098	96.16	6.38
FDRN [16]	0.7919	0.8830	94.69	5.70
CADDN (ours)	**0.8422**	**0.9139**	**96.27**	**5.57**

Table 2
Quantitative performance analysis on the CUHKMAP dataset. This table presents a comprehensive comparison of various shadow detection methods, including our proposed CADDN model, evaluated on the CUHKMAP dataset. Metrics include Intersection over Union (IoU), F1 Score, Overall Accuracy (OA), and Balanced Error Rate (BER).

Methods	IoU	F1	OA (%)	BER (%)
U-Net++ [50]	0.7564	0.8539	89.30	11.07
MaNet [52]	0.7587	0.8554	89.53	11.19
LinkNet [53]	0.7598	0.8563	89.51	10.97
HiFormer [60]	0.6944	0.8104	87.61	15.03
FSDNet [12]	0.7171	0.8261	87.15	13.39
MSASDNet [14]	0.7387	0.8421	88.46	12.05
OglaNet [61]	0.7483	0.8489	89.10	11.45
BDRAR [11]	0.7291	0.8355	88.15	12.74
DSC [15]	0.7387	0.8422	88.69	12.06
ECA [13]	0.7171	0.8273	87.44	12.71
FDRN [16]	0.7374	0.8410	88.33	12.18
CADDN (ours)	**0.7622**	**0.8581**	**89.64**	**10.88**

Table 3
Quantitative performance analysis on the SBU dataset. This table presents a comprehensive comparison of various shadow detection methods, including our proposed CADDN model, evaluated on the SBU dataset. Metrics include Intersection over Union (IoU), F1 Score, Overall Accuracy (OA), and Balanced Error Rate (BER).

Methods	IoU	F1	OA (%)	BER (%)
U-Net++ [50]	0.7053	0.7929	94.24	10.55
MaNet [52]	0.7220	0.8078	94.72	9.87
LinkNet [53]	0.7167	0.8021	94.56	10.45
HiFormer [60]	0.6965	0.7869	94.96	13.07
FSDNet [12]	0.6775	0.7691	94.02	12.66
MSASDNet [14]	0.6648	0.7574	93.38	12.71
OglaNet [61]	0.6880	0.7768	94.07	12.18
BDRAR [11]	0.7264	0.8094	95.24	10.70
DSC [15]	0.7295	0.8148	94.86	**8.70**
ECA [13]	0.6016	0.6983	91.78	16.76
FDRN [16]	0.6944	0.7862	93.69	8.95
CADDN (ours)	**0.7311**	**0.8154**	**95.27**	9.75

5. Discussion

To conduct a comprehensive analysis of the experiments, it is crucial to begin by comparing the quantitative differences among the considered methods across datasets. Therefore, let us start by discussing the results of each dataset in detail. In the case of AISD (Table 1), the proposed model (CADDN) achieves the best average results in all metrics, with an IoU of 0.8422, F1-score of 0.9139, OA of 96.27%, and BER of 5.57%. Following CADDN, MaNet and U-Net++ are the second and third best performing methods, obtaining lower yet still positive results. After them, OglaNet, ECA, LinkNet and MSASDNet also yield positive results, although with a slight performance decrease. On the other hand, the remaining methods are unable to produce satisfactory results in this data collection, being particularly negative for FSDNet and BDRAR. In CUHKMAP (Table 2), the best performance is also achieved by the proposed model, with quantitative values of 0.7622 (IoU), 0.8581 (F1-score), 89.64% (OA), and 10.88% (BER). The second and third best results are obtained by LinkNet and MaNet, which are followed by U-Net++, OglaNet, DSC and MSASDNet. However, DSC and MSASDNet experience a slight decrease in performance. Concerning the other methods, they tend to show a rather limited performance, which is particularly low for HiFormer.

Regarding the SBU dataset (Table 3), our proposed model is able to provide the best results for three of the four considered metrics, with an average IoU of 0.7311, F1-score of 0.8154, and OA of 95.27%. In this case, DSC and BDRAR exhibit the second and third best quantitative

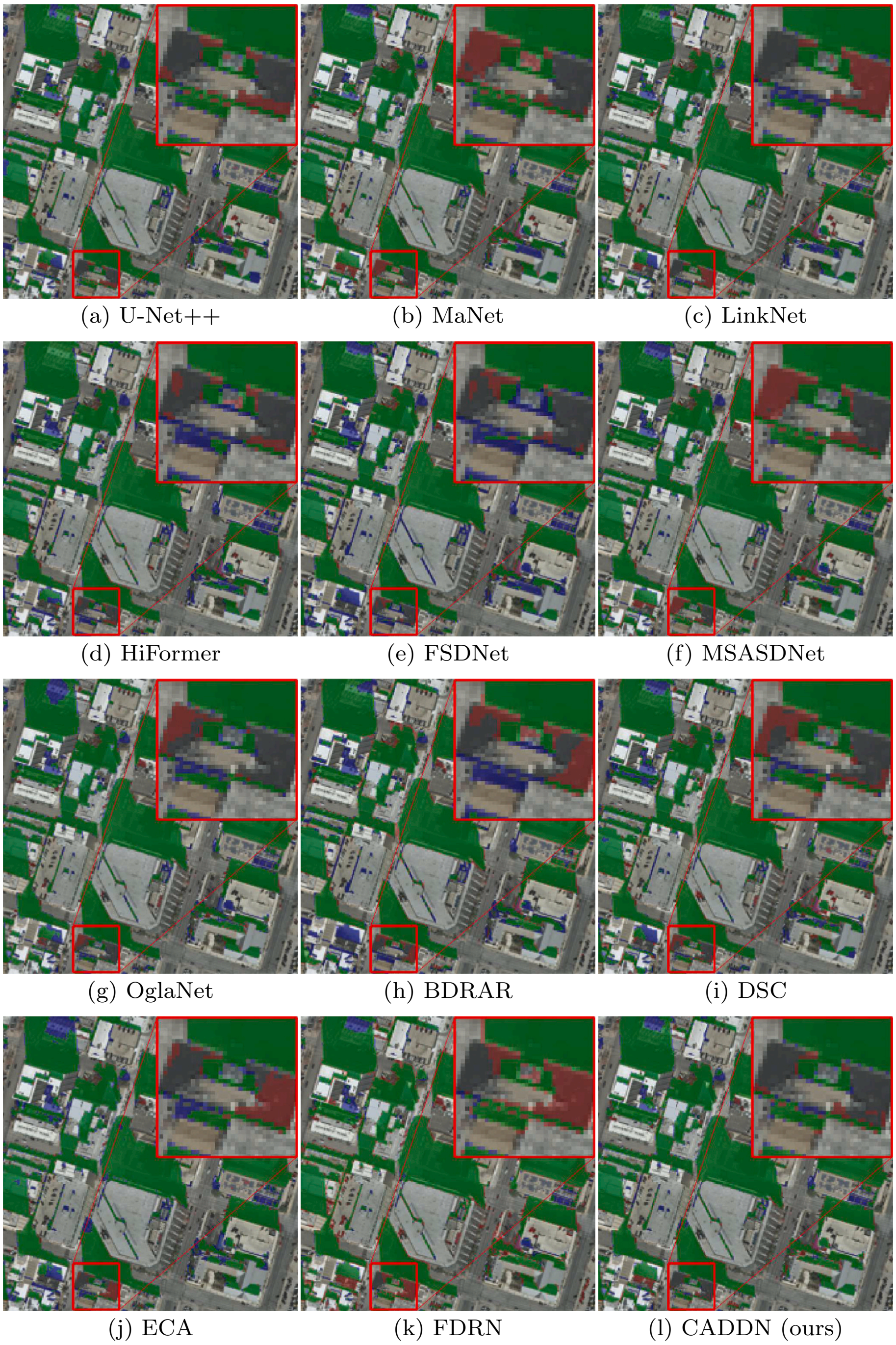


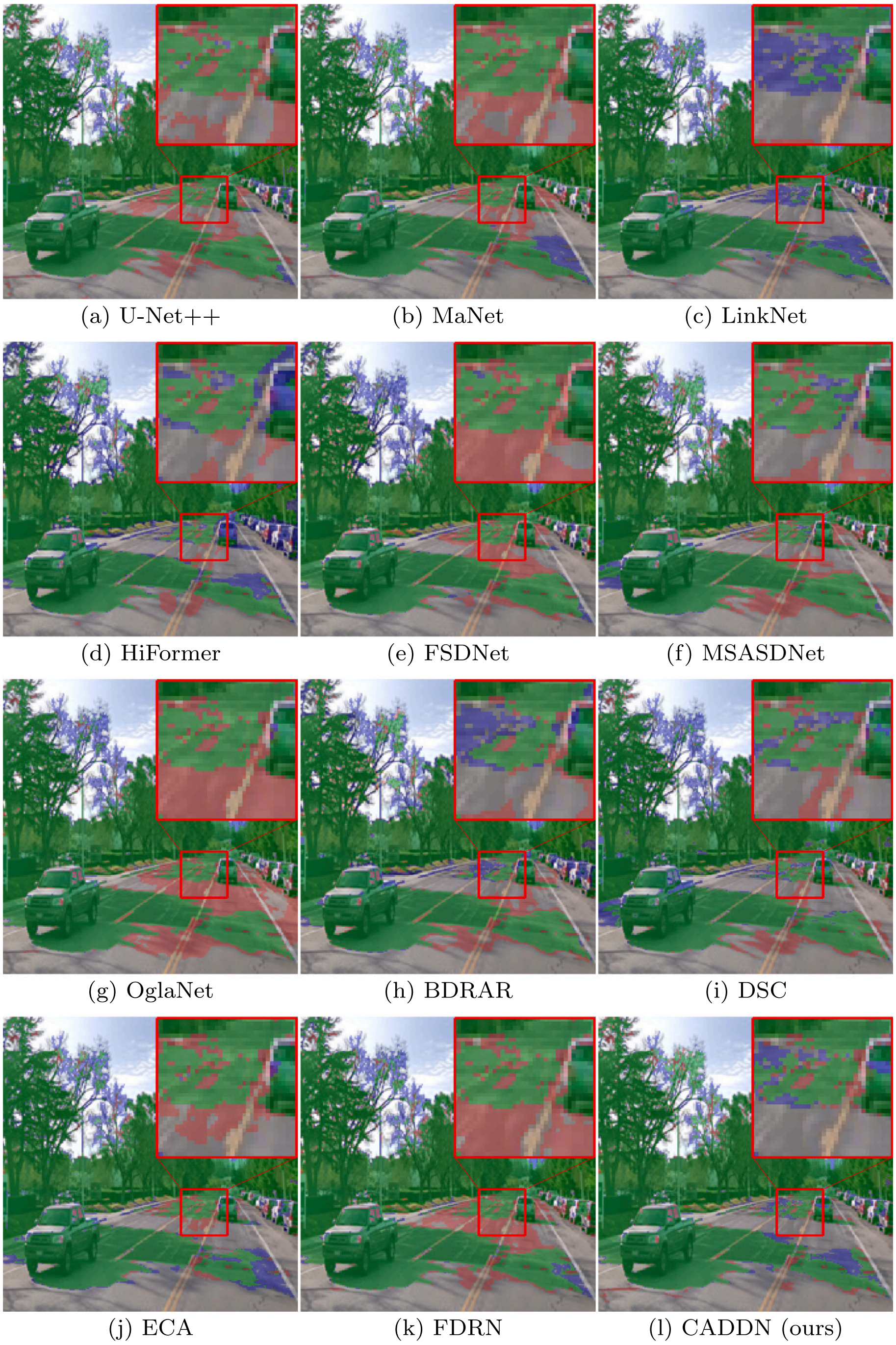
Fig. 7. Qualitative results of a sample test image from the AISD dataset. Note that true-positives are highlighted in green, false-positives in red and false-negatives in blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

results in these metrics by a small margin. Besides, OglaNet and MSASDNet seem to suffer a notable performance decrease with respect to other competitors. The remaining methods appear to achieve relatively limited performance in this collection, with ECA at the last position. Regarding the BER metric, it should be mentioned that DSC and FDRN obtained the best and second best quantitative results, with the proposed approach ranking third.

In light of these results, several important observations can be made on the basis of the nature of the considered methods and data. Overall, MaNet, LinkNet, and U-Net++ produce positive results across all the datasets, which reveals the robustness of these methods for detecting shadows in complex scenes, like those in AISD, CUHKMAP or SBU. Essentially, these models utilize symmetric encoder-decoder architectures that exploit skip connections to preserve contextual information from input data to output. By doing so, low-level features can be used to build higher-level characteristics or simply bypassed to directly detect the target. Note that this aspect can be very relevant in shadow detection, since shadows typically combine simple black regions that can be easily modeled by low-level features, with complex diffuse areas that logically require a better understanding of the objects in the scene. Although other models, such as MSASDNet or OglaNet, follow a similar idea using dense lateral connections, their performance has not been as good, particularly in CUHKMAP and SBU. This fact reveals that increasing the number of parameters in the decoder may not always lead to improved shadow detection, taking into account the experimental setup and training scheme considered in this work. Even more complex architectures, such as HiFormer, can exhibit relatively poor results in these types of scenario.

Another factor that appears to have a relevant impact on the results is the utilization of attention mechanisms. Although some of the evaluated models, such as U-Net++, obtain positive results without attention, while others with attention, such as HiFormer, can show limited performance, in general, it is possible to see that the models that integrate some form of attention demonstrate a greater consistency in performance across the datasets, as observed in the case of MaNet. Specifically, MaNet employs a transformer-based attention mechanism that weights the most relevant features after the encoder. The experimental results show that this type of configuration can be more effective for shadow detection than traditional spatial attentions, as in the case of MSASDNet, or applying attention in earlier stages, as in the case of HiFormer.

The positive results achieved by symmetric encoder-decoder architectures and the enhanced performance from implementing attention mechanisms at the decoder stage significantly endorse the design choices of our proposed network. CADDN not only employs these symmetric paths and late-stage attention in the shadow decoder segment, but also introduces other methodological advancements for shadow detection. Its dual-decoder configuration distinctively enables selective use of detailed image reconstruction features for shadow prediction. This is achieved through the integration of noisy skip connections, which enhance the features extracted by the image-decoder. Moreover, transfer blocks efficiently adapt these features for the shadow-decoder, accommodating the different types of data processed by each decoder

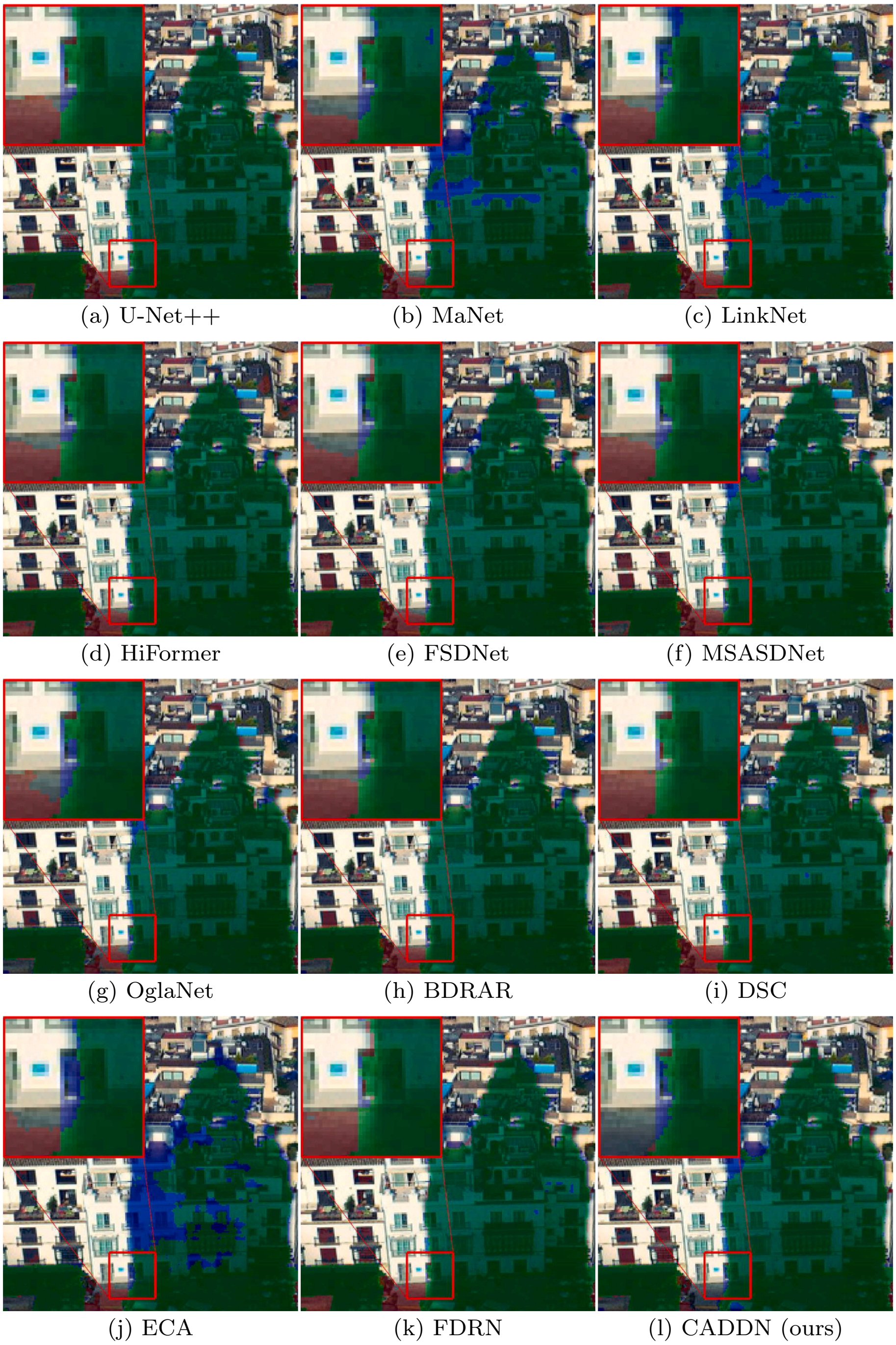

(a) U-Net++ (b) MaNet (c) LinkNet

(d) HiFormer (e) FSDNet (f) MSASDNet

(g) OglaNet (h) BDRAR (i) DSC

(j) ECA (k) FDRN (l) CADDN (ours)

Fig. 8. Qualitative results of a sample test image from the CUHKMAP dataset. Note that true-positives are highlighted in green, false-positives in red and false-negatives in blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



(a) U-Net++ (b) MaNet (c) LinkNet

(d) HiFormer (e) FSDNet (f) MSASDNet

(g) OglaNet (h) BDRAR (i) DSC

(j) ECA (k) FDRN (l) CADDN (ours)

Fig. 9. Qualitative results of a sample test image from the SBU dataset. Note that true-positives are highlighted in green, false-positives in red, and false-negatives in blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

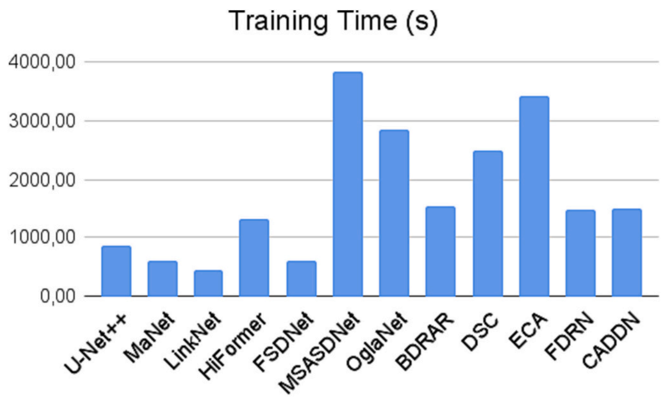


Fig. 10. Training time in seconds (s) for all the considered methods on the AISD dataset.

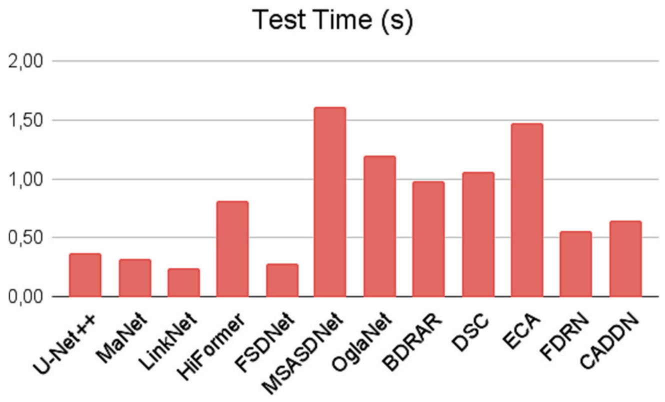


Fig. 11. Test time in seconds (s) for all the considered methods on the AISD dataset.

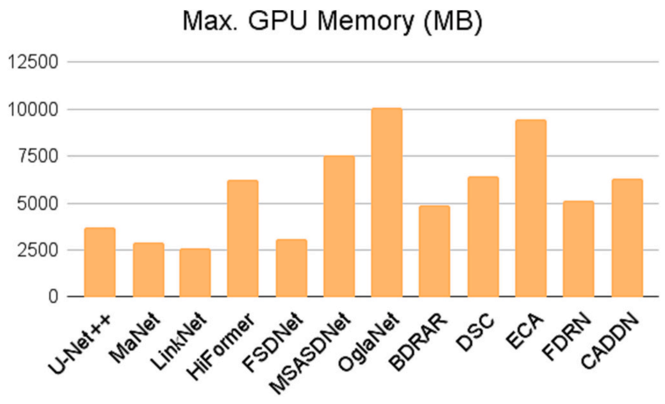


Fig. 12. Maximum GPU memory demand in megabytes (MB) for all the considered methods on the AISD dataset.

path. This unique approach of combining image and shadow decoders allows CADDN to excel across various datasets and evaluation metrics, demonstrating its effectiveness in shadow detection.

The empirical results from our comprehensive analysis consistently underscore the superior performance of the proposed network. Across different benchmark collections, such as AISD, CUHKMAP and SBU, CADDN consistently outperforms other methods in key metrics. This performance is a testament to the innovative features of CADDN, including its symmetric encoder-decoder paths, advanced attention mechanisms, and unique dual-decoder configuration. These novel features enable CADDN to achieve superior results in accurately detecting and differentiating shadows in a range of complex scenarios, as can be further validated in the presented visual results. Our qualitative analysis, illustrated in Figs. 7, 8 and 9, shows that CADDN consistently produces fewer false positives (marked red) and false negatives (marked in blue) compared to other models, especially in complex areas. For example, as seen in Fig. 8, CADDN outperforms other models in identifying diffuse shadows, such as those cast by a bare tree, which are commonly misclassified as asphalt. This level of precision is consistently demonstrated in other examples, as shown in Figs. 7 and 9.

6. Conclusions and future work

This paper presented a new deep learning-based shadow detection model (CADDN), which has been designed to deal with complex scenes. Specifically, the proposed model defines an innovative dual-decoder shape that includes two segments that work together to reconstruct both the input images and their corresponding shadow masks. In this way, fine-grained image reconstruction features can be transferred to support the shadow-decoder when necessary. Besides, the CADDN incorporates noisy skip connections and cross-attention to guarantee the quality of the transferred features. A new joint loss formulation is also defined to train the proposed shadow detection model based on the reconstructed images and the predicted masks. The experimental results, considering several benchmark datasets and shadow detection networks, demonstrate the competitive performance of the proposed approach.

The main conclusion that can be drawn from this research is that leveraging image reconstruction features is a viable approach to tackling the complexity of shadow detection. Challenges inherent to shadow detection are generally magnified in heterogeneous scenes, and hence the exploitation of fine-grained image reconstruction features can be a very useful tool in these scenarios. Despite the fact that the results obtained are certainly promising, there is always room for further improvements. As future work, we aim to expand this work in the following directions that could potentially lead to even better performances: exploring imbalanced loss formulations, additional decoder architectures, and adopting a semi-supervised scheme.

CRediT authorship contribution statement

Ruben Fernandez-Beltran: Conceptualization, Methodology, Software, Writing – original draft. **Angelica Guzman-Ponce:** Conceptualization, Methodology, Software, Writing – original draft. **Rafael Fernandez:** Conceptualization, Methodology, Software, Writing – original draft. **Jian Kang:** Data curation, Investigation, Writing – review & editing. **Ginés García-Mateos:** Funding acquisition, Project administration, Supervision, Writing – review & editing.

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.

Acknowledgements

This work was supported by the Postdoctoral Margarita Salas Fellowship MGS/2021/23(UP2021-021) from the European Union-NextGenerationEU funds, the Science and Technology Agency of the Region of Murcia (Fundación Séneca, Action Plan 2022) by grant 22130/PI/22, and the National Natural Science Foundation of China under Grant 62101371.

Appendix A. Parameter Sensitivity Analysis

Within the proposed model formulation, the η hyper-parameter modulates the amount of dropout noise that is injected to the encoder features in order to enhance the features uncovered by the image-decoder segment. In this section, we conduct a sensitivity analysis for this hyper-parameter to understand how η affects the performance of the proposed shadow detection network. Taking into account the experimental configuration described in Section 4.2, we test our CADDN model on the AIDS collection with the following dropout noise ratios $\eta = \{0.0, 0.1, 0.2, ..., 1.0\}$. Table A.1 reports the corresponding results based on IoU, F1 and OA metrics. As can be observed, introducing a moderate amount of noise into the encoder features tends to lead to certain performance improvements. These results can be attributed to the fact that introducing some noise into the encoder features forces the image-decoder to reconstruct such perturbations, which in turn enhances the ability of the network to uncover more accurate features along the image-decoder segment. However, these advantages seem to have a limit, since the performance of the model tends to decrease as the noise ratio increases beyond $\eta = 0.5$. In this case, excessive noise can disrupt the underlying patterns in the feature maps, making it harder for the decoder to reconstruct the original image and leading to an overall drop in performance. For the sake of generality, we suggest a default dropout noise ratio of $\eta = 0.1$.

Table A.1
Performance analysis across different noise levels. This table provides a sensitivity analysis of the proposed model with respect to different noise levels (η), assessing their impact on the Intersection over Union (IoU), F1 Score, and Overall Accuracy (OA) metrics. It illustrates the effect of varying degrees of dropout noise on the encoder features affect the shadow detection capabilities of our approach.

Noise Level (η)	IoU	F1	OA (%)
0.0	0.8413	0.9134	96.26
0.1	0.8424	0.9141	96.28
0.2	0.8413	0.9134	96.24
0.3	0.8410	0.9132	96.23
0.4	0.8420	0.9138	96.26
0.5	0.8420	0.9138	96.25
0.6	0.8402	0.9127	96.21
0.7	0.8406	0.9130	96.23
0.8	0.8409	0.9132	96.22
0.9	0.8387	0.9118	96.17
1.0	0.8356	0.9100	96.10

Appendix B. Ablation Study

Another important aspect to be analyzed is the contribution of the two components adopted by the proposed architecture, that is, the transfer blocks used between decoder segments and the considered cross-attention modules. To investigate this, we conducted an ablation study to compare our model with several simplified versions that remove all the transfer blocks (woTR) and the cross-attention modules (woCA). This comparison enables us to evaluate the real contribution of the proposed architecture compared to other configurations with the same elemental structure. We present the results of the ablation study in Table B.1, which includes the results based on the AISD dataset. The results show that the proposed CADDN consistently outperforms all the ablated versions, demonstrating the contribution of our newly designed architecture for shadow detection.

Table B.1
Ablation study for the proposed model. This table shows the performance metrics, including Intersection over Union (IoU), F1 Score, and Overall Accuracy (OA), for several simplified versions of the proposed architecture (CADDN). Specifically, CADDN-woTR represents CADDN without transfer blocks, CADDN-woCA without cross-attention modules and CADDN-none without transfer blocks neither cross-attention.

Version	Tr-Block	Cross-Att	IoU	F1	OA (%)
CADDN-none	No	No	0.8334	0.9087	96.02
CADDN-woCA	Yes	No	0.8368	0.9108	96.12
CADDN-woTR	No	Yes	0.8392	0.9121	96.20
CADDN	Yes	Yes	0.8423	0.9140	96.26

Appendix C. Trade-off analysis

Since the proposed loss is formulated according to two joint terms, i.e., image reconstruction ($\mathscr{L}_I$), and shadow matching ($\mathscr{L}_S$), this section analyzes the impact of the image reconstruction term on the overall performance of the proposed shadow detection model. Specifically, we vary the β trade-off parameter in Eq. (11) from 0.0 (null activation) to 1.0 (full activation) to weight the importance of the loss term $\mathscr{L}_I$. Table C.1 reports the analysis conducted on the AISD collection, using the experimental settings mentioned above. As is possible to observe, better average results can be achieved when non-zero values of β are considered, which reveals the positive contribution that the image-decoder features may have in the final shadow detection predictions. However, the performance differences among the non-null values tend to be small, indicating that once β is activated, the proposed model can be effective at different activation levels. Since the proposed joint loss is based on two bounded figures of merit that work in a similar value range, we set $\beta = 1.0$ as the default trade-off value to train our shadow detection model.

Table C.1
Impact of the β hyperparameter on the proposed model performance. This table details the sensitivity analysis of the β hyperparameter on the Intersection over Union (IoU), the F1 score and the Overall Accuracy (OA) of the proposed shadow detection model. The values of β range from 0 to 1, with each increment analyzed for its effect on performance metrics.

β	IoU	F1	OA (%)
0.0	0.8406	0.9130	96.23
0.1	0.8422	0.9139	96.26
0.2	0.8421	0.9139	96.27
0.3	0.8422	0.9140	96.26
0.4	0.8413	0.9134	96.26
0.5	0.8419	0.9137	96.25
0.6	0.8414	0.9134	96.24
0.7	0.8414	0.9134	96.25
0.8	0.8415	0.9135	96.25
0.9	0.8415	0.9135	96.25
1.0	0.8423	0.9140	96.26

References

[1] J.-F. Lalonde, A.A. Efros, S.G. Narasimhan, Estimating the natural illumination conditions from a single outdoor image, Int. J. Comput. Vis. 98 (2012) 123–145.

[2] X. Huang, G. Hua, J. Tumblin, L. Williams, What characterizes a shadow boundary under the sun and sky?, in: 2011 International conference on computer vision IEEE, 2011, pp. 898–905.

[3] W. Zhang, X. Zhao, J.-M. Morvan, L. Chen, Improving shadow suppression for illumination robust face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 41 (3) (2018) 611–624.

[4] R. Fernandez-Beltran, F. Pla, A. Plaza, Endmember extraction from hyperspectral imagery based on probabilistic tensor moments, IEEE Geosci. Remote Sens. Lett. 17 (12) (2020) 2120–2124.

[5] L. Yang, P. Bi, H. Tang, F. Zhang, Z. Wang, Improving vegetation segmentation with shadow effects based on double input networks using polarization images, Comput. Electron. Agric. 199 (2022) 107123.

[6] N. Al-Najdawi, H.E. Bez, J. Singhai, E.A. Edirisinghe, A survey of cast shadow detection algorithms, Pattern Recogn. Lett. 33 (6) (2012) 752–764.

[7] A. Sanin, C. Sanderson, B.C. Lovell, Shadow detection: a survey and comparative evaluation of recent methods, Pattern Recogn. 45 (4) (2012) 1684–1695.

[8] Z. Li, H. Shen, Q. Weng, Y. Zhang, P. Dou, L. Zhang, Cloud and cloud shadow detection for optical satellite imagery: features, algorithms, validation, and prospects, ISPRS Journal of Photogrammetry and Remote Sensing 188 (2022) 89–108.

[9] I. Huerta, M.B. Holte, T.B. Moeslund, J. Gonzàlez, Chromatic shadow detection and tracking for moving foreground segmentation, Image Vis. Comput. 41 (2015) 42–53.

[10] J. Zhu, K.G. Samuel, S.Z. Masood, M.F. Tappen, Learning to recognize shadows in monochromatic natural images, in: 2010 IEEE Computer Society Conference on Computer Vision and pattern recognition, IEEE, 2010, pp. 223–230.

[11] L. Zhu, Z. Deng, X. Hu, C.-W. Fu, X. Xu, J. Qin, P.-A. Heng, Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 121–136.

[12] X. Hu, T. Wang, C.-W. Fu, Y. Jiang, Q. Wang, P.-A. Heng, Revisiting shadow detection: a new benchmark dataset for complex world, IEEE Trans. Image Process. 30 (2021) 1925–1934.

[13] X. Fang, X. He, L. Wang, J. Shen, Robust shadow detection by exploring effective shadow contexts, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 2927–2935.

[14] D. Liu, J. Zhang, Y. Wu, Y. Zhang, A shadow detection algorithm based on multiscale spatial attention mechanism for aerial remote sensing images, IEEE Geosci. Remote Sens. Lett. 19 (2021) 1–5.

[15] X. Hu, C.-W. Fu, L. Zhu, J. Qin, P.-A. Heng, Direction-aware spatial context features for shadow detection and removal, IEEE Trans. Pattern Anal. Mach. Intell. 42 (11) (2019) 2795–2808.

[16] L. Zhu, K. Xu, Z. Ke, R.W. Lau, Mitigating intensity bias in shadow detection via feature decomposition and reweighting, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 4702–4711.

[17] T. Zhou, H. Fu, C. Sun, S. Wang, Shadow detection and compensation from remote sensing images under complex urban conditions, Remote Sens. (Basel) 13 (4) (2021) 699.

[18] T. Wang, X. Hu, Q. Wang, P.-A. Heng, C.-W. Fu, Instance shadow detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1880–1889.

[19] R. Vinuesa, H. Azizpour, I. Leite, M. Balaam, V. Dignum, S. Domisch, A. Felländer, S.D. Langhans, M. Tegmark, F. Fuso Nerini, The role of artificial intelligence in achieving the sustainable development goals, Nat. Commun. 11 (1) (2020) 1–10.

[20] J.-F. Lalonde, A.A. Efros, S.G. Narasimhan, Detecting ground shadows in outdoor consumer photographs, in: Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5–11, 2010, Proceedings, Part II 11, Springer, 2010, pp. 322–335.

[21] G. Zhou, Y. Tang, W. Zhang, W. Liu, Y. Jiang, E. Gao, Q. Zhu, Y. Bai, Shadow detection on high-resolution digital orthophoto map (dom) using semantic matching, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–20.

[22] N. Bansal, N. Aggarwal, Deep learning based shadow detection in images, in: Proceedings of 2nd International Conference on Communication, Computing and Networking: ICCCN 2018, NITTTR Chandigarh, India, Springer, 2019, pp. 375–382.

[23] S.H. Khan, M. Bennamoun, F. Sohel, R. Togneri, Automatic shadow detection and removal from a single image, IEEE Trans. Pattern Anal. Mach. Intell. 38 (3) (2015) 431–446.

[24] T.F.Y. Vicente, L. Hou, C.-P. Yu, M. Hoai, D. Samaras, Large-scale training of shadow detectors with noisily-annotated shadow examples, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VI 14, Springer, 2016, pp. 816–832.

[25] S. Hosseinzadeh, M. Shakeri, H. Zhang, Fast shadow detection from a single image using a patched convolutional neural network, in: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2018, pp. 3124–3129.

[26] L. Jie, H. Zhang, When sam meets shadow detection, arXiv preprint (2023) arXiv: 2305.11513.

[27] L. Jiao, M. Zheng, P. Tang, Z. Zhang, Towards edge-precise cloud and shadow detection on the gaofen-1 dataset: a visual, comprehensive investigation, Remote Sens. (Basel) 15 (4) (2023) 906.

[28] J. Zhang, X. Shi, C. Zheng, J. Wu, Y. Li, Mrpfa-net for shadow detection in remote-sensing images, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–11, https://doi.org/10.1109/TGRS.2023.3282967.

[29] A. Kumar, Seat-yolo: a squeeze-excite and spatial attentive you only look once architecture for shadow detection, Optik 273 (2023) 170513.

[30] H. Zhou, J. Yi, Ffsdf: an improved fast face shadow detection framework based on channel spatial attention enhancement, Journal of King Saud University-Computer and Information Sciences 35 (9) (2023) 101766.

[31] L. Liu, J. Prost, L. Zhu, N. Papadakis, P. Liò, C.-B. Schönlieb, A.I. Aviles-Rivero, Scotch and soda: A transformer video shadow detection framework, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 10449–10458.

[32] M.K. Yücel, V. Dimaridou, B. Manganelli, M. Ozay, A. Drosou, A. Saa-Garriga, Lra&ldra: Rethinking residual predictions for efficient shadow detection and removal, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 4925–4935.

[33] R. Cong, Y. Guan, J. Chen, W. Zhang, Y. Zhao, S. Kwong, Sddnet: Style-guided dual-layer disentanglement network for shadow detection, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 1202–1211.

[34] X. Zhang, J. Li, C. Yang, C. Zhang, Cifnet: context information fusion network for cloud and cloud shadow detection in optical remote sensing imagery, J. Appl. Remote. Sens. 17 (1) (2023) 016506.

[35] C. Zhang, L. Weng, L. Ding, M. Xia, H. Lin, Crsnet: cloud and cloud shadow refinement segmentation networks for remote sensing imagery, Remote Sens. (Basel) 15 (6) (2023) 1664.

[36] J. Feng, J. Liu, Z. Gu, W. Zheng, Oamsfnet: orientation-aware and multi-scale feature fusion network for shadow detection in remote sensing images via pseudo shadow, Int. J. Remote Sens. 44 (17) (2023) 5473–5495.

[37] J. Chen, G. Xing, J. Liao, H. Wei, Y. Liu, Boundary-aware shadow detection via mask decoupling and feature correction, in: 2023 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2023, pp. 150–155.

[38] W. Wu, K. Zhou, X.-D. Chen, Single image shadow detection via uncertainty analysis and gcn-based refinement strategy, J. Vis. Commun. Image Represent. 82 (2022) 103397.

[39] W. Wu, K. Zhou, X.-D. Chen, J.-H. Yong, Light-weight shadow detection via gcn-based annotation strategy and knowledge distillation, Comput. Vis. Image Underst. 216 (2022) 103341.

[40] Z. Zhang, W. Shen, L. Xia, Y. Lin, S. Shang, W. Hong, Video Sar moving target shadow detection based on intensity information and neighborhood similarity, Remote Sens. (Basel) 15 (7) (2023) 1859.

[41] J.M.J. Valanarasu, V.M. Patel, Fine-context shadow detection using shadow removal, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 1705–1714.

[42] L. Zhang, C. Long, X. Zhang, C. Xiao, Exploiting residual and illumination with gans for shadow detection and shadow removal, ACM Trans. Multimed. Comput. Commun. Appl. 19 (3) (2023) 1–22.

[43] X. Zhang, Y. Zhao, C. Gu, C. Lu, S. Zhu, Spa-former: An effective and lightweight transformer for image shadow removal, in: 2023 International Joint Conference on Neural Networks (IJCNN), IEEE, 2023, pp. 1–8.

[44] L. Guo, C. Wang, W. Yang, S. Huang, Y. Wang, H. Pfister, B. Wen, Shadowdiffusion: When degradation prior meets diffusion model for shadow removal, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 14049–14058.

[45] W.-J. Ahn, G. Kang, H.-D. Choi, M.-T. Lim, Domain adaptation for complex shadow removal with shadow transformer network, Neurocomputing 552 (2023) 126559.

[46] Y. Xu, M. Lin, H. Yang, F. Chao, R. Ji, Shadow-aware dynamic convolution for shadow removal, Pattern Recogn. 146 (2024) 109969.

[47] H. Le, D. Samaras, From shadow segmentation to shadow removal, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16, Springer, 2020, pp. 264–281.

[48] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18, Springer, 2015, pp. 234–241.

[49] L. Jiao, L. Huo, C. Hu, P. Tang, Refined unet: Unet-based refinement network for cloud and shadow precise segmentation, Remote Sens. (Basel) 12 (12) (2020) 2001.

[50] Z. Zhou, M.M. Rahman Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested u-net architecture for medical image segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4, Springer, 2018, pp. 3–11.

[51] H. Li, P. Xiong, J. An, L. Wang, Pyramid attention network for semantic segmentation, Proceedings of the British Machine Vision Conference (2018) 1–13.

[52] T. Fan, G. Wang, Y. Li, H. Wang, Ma-net: a multi-scale attention network for liver and tumor segmentation, IEEE Access 8 (2020) 179656–179665.

[53] A. Chaurasia, E. Culurciello, Linknet: Exploiting encoder representations for efficient semantic segmentation, in: 2017 IEEE visual communications and image processing (VCIP), IEEE, 2017, pp. 1–4.

[54] S. Seferbekov, V. Iglovikov, A. Buslaev, A. Shvets, Feature pyramid network for multi-class land segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 272–275.

[55] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2881–2890.

[56] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 6 (2017) 1–14.

[57] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 801–818.

[58] Y. Liu, Y. Zhang, Y. Wang, F. Hou, J. Yuan, J. Tian, Y. Zhang, Z. Shi, J. Fan, Z. He, A survey of visual transformers, in: IEEE Transactions on Neural Networks and Learning Systems, 2023.

[59] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, Segformer: simple and efficient design for semantic segmentation with transformers, Adv. Neural Inf. Proces. Syst. 34 (2021) 12077–12090.

[60] M. Heidari, A. Kazerouni, M. Soltany, R. Azad, E.K. Aghdam, J. Cohen-Adad, D. Merhof, Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 6202–6212.

[61] Y. Xie, D. Feng, H. Chen, Z. Liao, J. Zhu, C. Li, S.W. Baik, An omni-scale global–local aware network for shadow extraction in remote sensing imagery, ISPRS Journal of Photogrammetry and Remote Sensing 193 (2022) 29–44.

[62] Y. Zhu, X. Fu, C. Cao, X. Wang, Q. Sun, Z.-J. Zha, Single image shadow detection via complementary mechanism, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 6717–6726.

[63] W. Wu, X.-D. Chen, W. Yang, J.-H. Yong, Exploring better target for shadow detection, Knowl.-Based Syst. 273 (2023) 110614.

[64] L. Jie, H. Zhang, Rmlanet: random multi-level attention network for shadow detection and removal, IEEE Trans. Circuits Syst. Video Technol. 33 (2023) 7819–7831.

[65] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[66] Z. Fan, Y. Liu, M. Xia, J. Hou, F. Yan, Q. Zang, Resat-unet: a u-shaped network using resnet and attention module for image segmentation of urban buildings, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 16 (2023) 2094–2111.

[67] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, W. Liu, Ccnet: Criss-cross attention for semantic segmentation, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 603–612.

[68] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.

[69] G. Zhai, X. Min, Perceptual image quality assessment: a survey, Science China, Inform. Sci. 63 (2020) 1–52.

[70] D. Brunet, E.R. Vrscay, Z. Wang, On the mathematical properties of the structural similarity index, IEEE Trans. Image Process. 21 (4) (2011) 1488–1499.

[71] S. Jadon, A survey of loss functions for semantic segmentation, in: 2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB), IEEE, 2020, pp. 1–7.

[72] C.H. Sudre, W. Li, T. Vercauteren, S. Ourselin, M. Jorge Cardoso, Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3, Springer, 2017, pp. 240–248.

[73] T. Wang, X. Hu, P.-A. Heng, C.-W. Fu, Instance shadow detection with a single-stage detector, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2023) 3259–3273.

[74] S. Luo, H. Li, H. Shen, Deeply supervised convolutional neural network for shadow detection based on a novel aerial shadow imagery dataset, ISPRS J. Photogramm. Remote Sens. 167 (2020) 443–457.

[75] E. Maggiori, Y. Tarabalka, G. Charpiat, P. Alliez, Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark, 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE, 2017, pp. 3226–3229.

[76] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13, Springer, 2014, pp. 740–755.