applied sciences

MDPI

Review

Deep Learning Methods in Image Matting: A Survey

Lingtao Huang [1,†], Xipeng Liu [1,†], Xuelin Wang [2], Jiangqi Li [2] and Benying Tan [2,*]

[1] School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, China; hlt1282075034@gmail.com (L.H.); xipengliu2002@gmail.com (X.L.)
[2] School of Artificial Intelligence, Guilin University of Electronic Technology, Guilin 541004, China; 2001620218@mails.guet.edu.cn (X.W.); 626231846@mails.guet.edu.cn (J.L.)
* Correspondence: by-tan@guet.edu.cn
† These authors contributed equally to this work.

Abstract: Image matting is a fundamental technique used to extract a fine foreground image from a given image by estimating the opacity values of each pixel. It is one of the key techniques in image processing and has a wide range of applications in practical scenarios, such as in image and video editing. Deep learning has demonstrated outstanding performance in various image processing tasks, making it a popular research topic. In recent years, image matting methods based on deep learning have gained significant attention due to their superior performance. Therefore, this article presents a comprehensive overview of the deep learning-based image matting algorithms that have been proposed in recent years. This paper initially introduces frequently used datasets and their production methods, along with the basic principles of traditional image matting techniques. We then analyze deep learning-based matting algorithms in detail and introduce commonly used image matting evaluation metrics. Additionally, this paper discusses the application scenarios of image matting, conducts experiments to illustrate the limitations of current image matting methods, and outlines potential future research directions in this field. Overall, this paper can serve as a valuable reference for researchers that are interested in image matting.

Keywords: image matting; deep learning; trimap; image composition; alpha matte

Citation: Huang, L.; Liu, X.; Wang, X.; Li, J.; Tan, B. Deep Learning Methods in Image Matting: A Survey. *Appl. Sci.* **2023**, *13*, 6512. https://doi.org/10.3390/app13116512

Academic Editor: Yu-Dong Zhang and Zhonghua Sun

Received: 21 March 2023
Revised: 17 May 2023
Accepted: 23 May 2023
Published: 26 May 2023

1. Introduction

1.1. Image Matting

Image matting [1] is a technique that is used for extracting the fine foreground of images, which is a computer vision task that has a wide range of application scenarios. The foreground of an image may include entities such as humans or animals, with delicate and complex edges such as hair as well as transparent objects such as glass, light bulbs, or water. These elements can be difficult for computers to accurately recognize. Image matting calculates the opacity of each pixel in an input image to obtain the alpha matte, which allows for the separation of the foreground from the background. The foreground can be composited with any background image to obtain a new image, as shown in Figure 1.

Chroma key matting [2] is a classic image matting technique that can be used to obtain the foreground from a solid background by adjusting the colors of pixels in the background to make them transparent. This technique requires a special shooting environment, which is considerably limited in practical applications. Therefore, researchers have focused on extracting alpha mattes from images that have natural scenes as the background. However, image matting in natural scenes loosens the constraint of setting a solid color background, which leads to a decline in alpha matte accuracy. Mathematically, the problem of image matting can be expressed as follows:

$$I_i = \alpha_i F_i + (1 - \alpha_i) B_i \quad (1)$$

where I_i denotes the RGB value at pixel i of the input image, α_i denotes the opacity value at pixel i ranging between 0 and 1, F_i denotes the RGB value at pixel i in the foreground, and B_i denotes the RGB value at pixel i in the background.

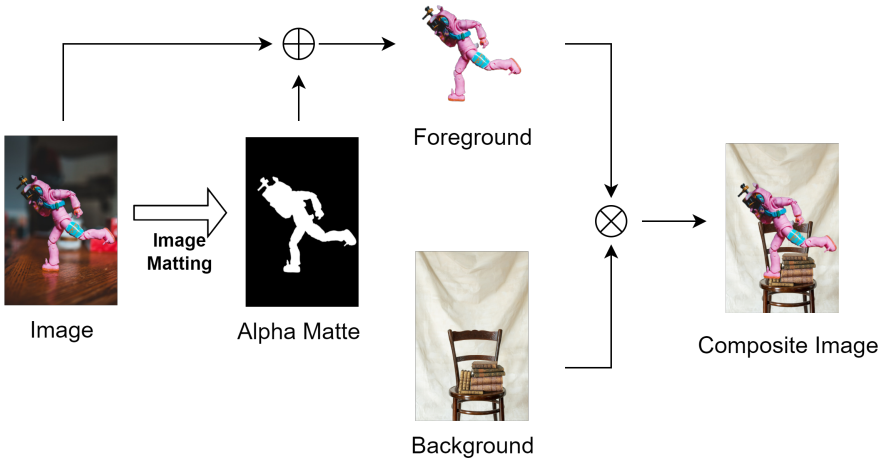

Figure 1. Specific process of image matting. $\oplus$ indicates the foreground extraction operation, and $\otimes$ indicates the image composition operation.

In image matting, the input image can be represented as a linear combination of foreground and background, with each pixel having only three known variables (RGB values) but seven unknown variables to be solved, as shown in Formula (1). Moreover, the definition of the foreground in an image is not precise and varies depending on the intended use. Consequently, image matting is a highly ill-posed problem that typically necessitates the use of auxiliary inputs such as trimaps to provide additional information.

1.2. Trimap

The trimap technique was proposed by Sun et al. [3] in 2004. A trimap is a mask that contains a foreground region, a background region, and an unknown region, and the regions are represented by α values of 1, 0, and 0.5, respectively. Figure 2 shows the image and corresponding trimap. The trimap can either be manually provided by users or automatically produced during network computations. When using the trimap as the auxiliary input, the known foreground and background regions provide considerable prior information, which helps to narrow down the solution space for unknown regions. Researchers can design relevant algorithms by using the information of these known regions.

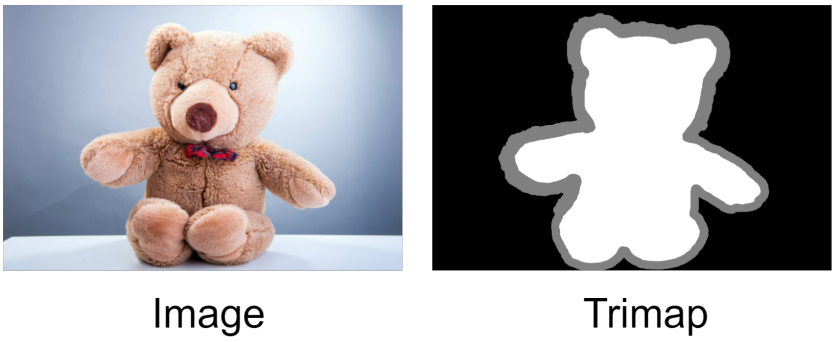

Figure 2. The image and its corresponding trimap are presented, where the foreground, background, and unknown regions of the trimap are denoted by white, black, and gray colors, respectively.

1.3. Distinguishing Image Matting from Image Semantic Segmentation

The results generated by image matting may appear similar to those of semantic segmentation; however, in reality, they are fundamentally different techniques. Semantic

segmentation is a classification task that extracts the semantic information in the input image and then classifies the pixels individually to obtain the semantic mask of the input image. When semantic segmentation only segments the foreground and background, the binary nature of segmentation leads to a strict boundary near the foreground edge. Image matting is a regression task that involves estimating the opacity of each pixel in an input image, which results in the extraction of the foreground via the alpha matte. A comparison of the results of semantic segmentation and image matting is shown in Figure 3.

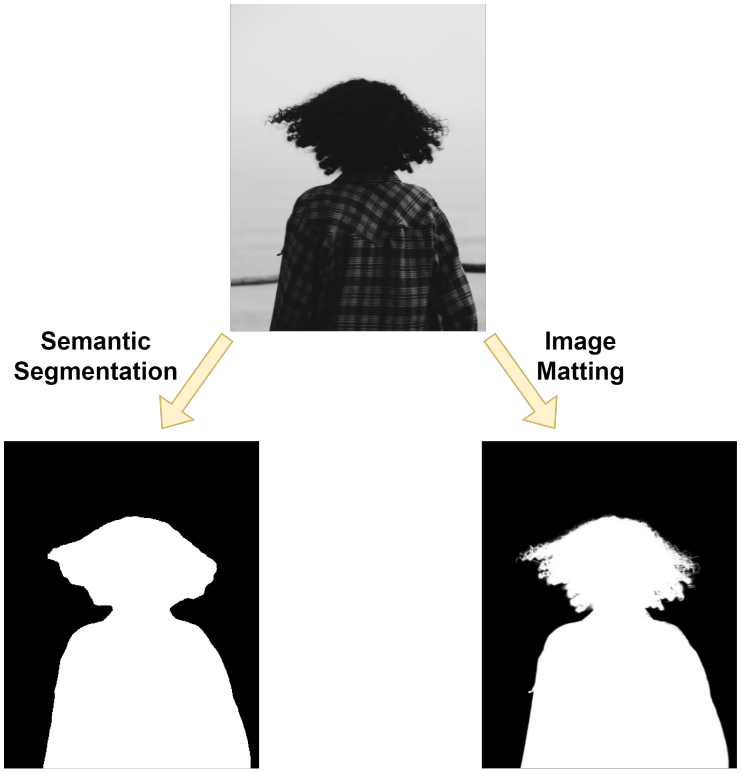

Figure 3. Comparison of the image semantic segmentation and image matting results.

1.4. Classification of Image Matting Methods

Over the years of research and development, researchers have designed a series of effective algorithms for various application scenarios of image matting; these algorithms can be categorized into three groups: sampling-based, propagation-based, and learning-based methods. Sampling-based algorithms predict the opacity of the unknown region by collecting a series of pixels or pixel blocks from the known regions of the trimap. Propagation-based algorithms typically establish connections between adjacent pixels and then use an optimization strategy to propagate the opacity information of the known regions to the unknown regions in order to predict the opacity of each pixel in the unknown region. Learning-based algorithms learn the features of the image by using a considerable amount of data and use these features to predict the opacity. As deep learning algorithms have already been applied to various visual tasks and have completely surpassed traditional learning-based algorithms, they are gradually being introduced into image matting.

1.5. Contributions

Traditional image matting algorithms (non-deep learning algorithms) have already been comprehensively reviewed [4–7], but there is still no comprehensive review of deep learning-based image matting methods. Thus, this article provides an extensive overview of deep learning-based image matting methods. The timeline of the deep learning algorithms

mentioned in this paper is shown in Figure 4. The main contributions of this article are as follows:

- We present a comprehensive overview of deep learning-based image matting approaches;
- We extensively introduce and analyze innovative deep learning-based image matting algorithms that have been developed in recent years and highlight their advantages;
- We outline the dataset, evaluation metrics, application scenarios, challenges, and potential research directions in image matting;

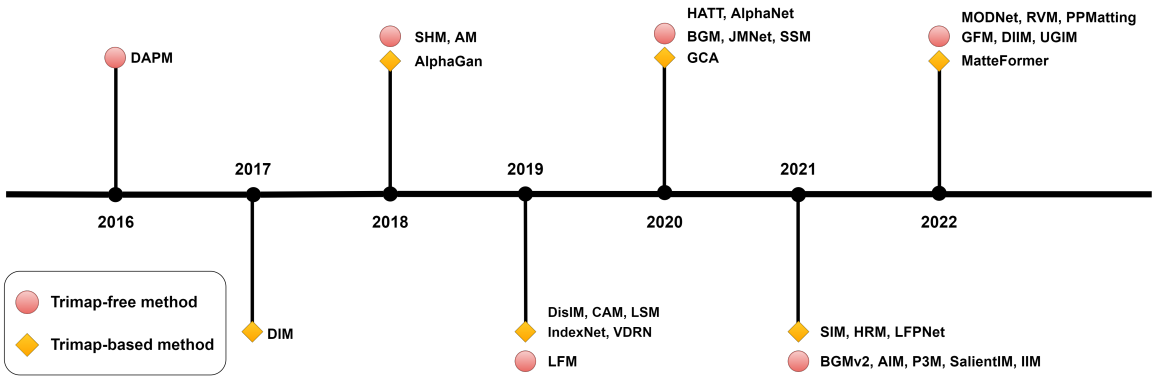

Figure 4. The timeline of deep learning-based matting methods is presented with the methods being listed chronologically according to the year they were proposed. Trimap-based methods are denoted by circles, whereas trimap-free methods are denoted by diamonds. (Related methods and corresponding references: DAPM [8], DIM [9], SHM [10], AM [11], AlphaGan [12], DisIM [13], CAM [14], LSM [15], IndexNet [16], VDRN [17], LFM [18], HATT [19], AlphaNet [20], BGM [21], JMNet [22], SSM [23], GCA [24], SIM [25], HRM [26], LFPNet [27], BGMV2 [28], AIM [29], P3M [30], SalientIM [31], IIM [32], MODNet [33], RVM [34], PPMatting [35], GFM [36], DIIM [37], UGIM [38], MatteFormer [39]).

1.6. Organization of the Remaining Sections

The subsequent sections of this article are arranged in the following order. Section 2 discusses some well-known image matting datasets and introduces the method of creating a dataset. Section 3 introduces the fundamental principles of various traditional methods and analyzes their advantages and disadvantages. Section 4 provides an overview of the existing end-to-end deep learning matting algorithms. Section 5 introduces some commonly used quality evaluation metrics of image matting. Section 6 introduces the application scenarios of image matching. Section 7 presents the current challenges and potential research directions in image matting. Section 8 provides a summary and conclusion.

2. Datasets

Rhemann et al. [40] created a benchmark comprising 27 training images and 8 test images that can be utilized to evaluate the accuracy of online image matting algorithms in terms of alpha matte quality. The images and corresponding trimaps in this dataset have two different resolution sizes, small and large, which can be used for different application scenarios based on specific requirements. This dataset is the initial benchmark that is specifically designed for image matting, which offers a reliable benchmark for assessing the effectiveness of image matting approaches.

Xu et al. [9] created a large-scale dataset named Composition-1K, which contains 49,300 training images and 1000 test images. This dataset was created by manually extracting the foreground from the image using image processing software and then compositing it onto a new background image. This is the inaugural large-scale image matting dataset, which furnishes an ample quantum of data for the training of neural networks.

Shen et al. [8] created a portrait dataset containing 2000 images. The authors used advanced image matting algorithms to obtain the alpha mattes of each image and then

constructed the dataset by manually selecting high-quality alpha mattes. The alpha mattes of each image contained detailed information that was necessary for training and validation. However, noise is inevitably introduced into the alpha mattes of this dataset during algorithm execution.

Sun et al. [25] produced a semantic dataset with extensive coverage of image matting patterns, taking into account the balance of data across various semantic categories. This balance ensures that images containing foreground objects in each category are almost equal in quantity.

Chen et al. [10] created a large-scale and high-quality portrait matting dataset containing 52,511 images. This dataset includes the foregrounds of humans carrying accessories such as mobile phones and handbags. After recruiting volunteers and carefully screening, the researchers obtained 35,311 foregrounds with high-quality alpha mattes. Additionally, they also selected 202 human foregrounds from the Composition-1K dataset [9]. Background images without humans were selected from the COCO dataset [41] and from the Internet, and they were then composited with the foregrounds.

Qiao et al. [19] created a dataset called Distinctions-646, which is a large-scale matting dataset consisting of 59,600 training images and 1000 test images. This dataset contains a total of 646 foregrounds composited with different backgrounds.

Lin et al. [28] created two datasets named VideoMatte240K and PhotoMatte13K, respectively. VideoMatte240K was created by collecting 484 high-definition green screen videos at 4K resolution. Chroma key software was then used to generate 240,709 frames of alpha mattes and foregrounds. The dataset includes people with different outfits and postures. It is noteworthy that this is the initial video matting dataset consisting of consecutive frames, which is distinct from static images and offers immense potential for future research on video and motion information. PhotoMatte13K comprises 13,665 images taken in a studio setting against a green screen background. Although the dataset offers a limited number of people poses, it boasts a high resolution, which allows it to capture finer details.

Li et al. [29] developed the AIM-500 benchmark, which comprises a diverse set of natural images of different types. This benchmark is equipped with high-quality alpha mattes, which provides a means for evaluating the generalization capability of algorithms.

Ke et al. [33] created the PPM-100 benchmark, which contains 100 finely annotated portraits with different backgrounds. The benchmark defines several classification rules to balance the sample types, such as the inclusion of an entire human body, existence of background blur, and whether the person is holding objects.

Li ei al. [36] have created two large-scale natural image datasets, namely AM-2K and BG-20K. AM-2K contains 2000 high-quality animal images along with a finely annotated alpha matte. YOLO-V3 [42] was utilized by the authors to construct BG-20K, ensuring the absence of salient objects in each image. BG-20K encompasses 20,000 high-resolution background images, which can improve the quality of matting and composition datasets by avoiding cluttered backgrounds that may affect model training.

Li et al. [30] developed a large-scale anonymized portrait dataset named P3M-10K with a primary focus on portrait privacy issues, setting it apart from other portrait matting datasets. This dataset includes 10,000 high-resolution portrait images in which faces are blurred and alpha mattes are of high quality. As the privacy of datasets has gained increasing attention, this dataset provides a solid foundation for future research on the impact of privacy-preserving training on portrait matting.

Building Matting Datasets

The primary challenge in creating an image matting dataset is obtaining high-quality alpha mattes for images. The alpha matte labels the opacity value of each pixel, which cannot be directly discerned by the human eye. Consequently, opacity values are obtained through the following methods: (1) using chroma key: a scene is initially set up with the foreground and camera fixed while capturing the natural image. Subsequently, a screen is placed between the natural background and the foreground and is set to alternate be-

tween red, green, and blue, capturing images of the foreground against different color backgrounds. The chroma key is subsequently applied to obtain the alpha mattes from the images, and the final alpha matte is computed by taking the average of these alpha mattes. High-quality alpha mattes can be obtained from different natural scenes by repeating this procedure; (2) using image processing softwares (e.g., Photoshop [43]): the alpha matte obtained by this method is rough and influenced by human subjectivity; and (3) using current state-of-the-art image matting algorithms: this approach enables the rapid attainment of high-quality outcomes. However, the performance of the algorithm constrains the quality of the obtained alpha mattes, and the quality of the alpha mattes obtained from different types of natural images varies significantly.

Additionally, trimaps should be included in an image matting dataset alongside alpha mattes. There are two main methods of generating trimaps: (1) manually annotating: the trimaps generated through this approach demonstrate high accuracy, but the process is exceedingly costly; and (2) using alpha mattes: this method utilizes image processing techniques, such as dilation and erosion, to obtain trimaps from alpha mattes, which is a low-cost approach. However, the trimaps generated by this method may have a lower accuracy compared with those generated by manually annotating, as shown in Figure 5.

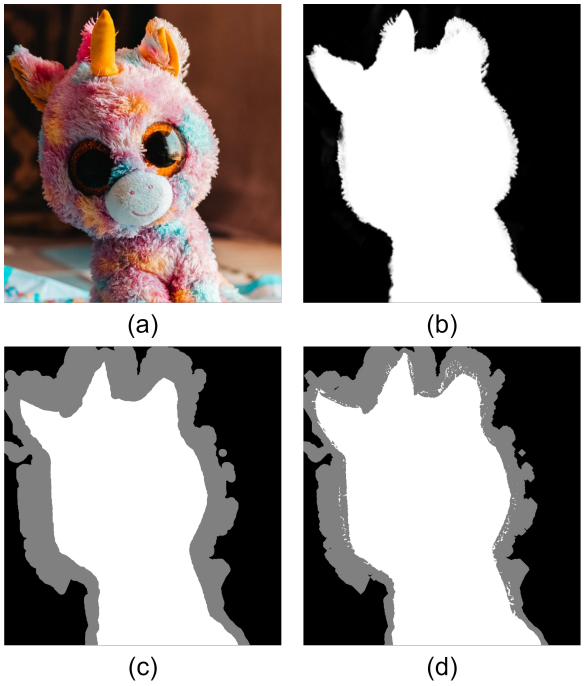

Figure 5. Image matting process: (**a**) original image, (**b**) corresponding alpha matte, (**c**) manual trimap, and (**d**) trimap automatically generated by expansion and corrosion. It is observed that (**d**) is significantly rougher than (**c**).

Due to the high cost of obtaining alpha mattes, some datasets obtain the foreground by using the alpha matte and then compositing it with different backgrounds to generate new images and expand the dataset. Although this method can quickly increase the size of the dataset, datasets that contain a large number of composited images with compositing artefacts may affect the generalization ability of models in applications [9,10,19].

3. Traditional Methods

A large number of concepts and techniques in deep learning methods are derived from traditional methods. Therefore, we introduce the fundamental principles of traditional methods before delving into the remarkable deep learning methods that have been proposed in recent years. Traditional methods primarily rely on low-level image features, such as color and brightness differences between the foreground and background, as well

as handcrafted features to extract alpha mattes; there are the three main categories of traditional image matting methods: the sampling-based method, the propagation-based method (i.e., affinity-based), and the hybrid method based on both the sampling-based and propagation-based methods.

3.1. Sampling-Based Methods

Sampling-based methods [44–49] sample a series of pixel pairs from known regions of the trimap before predicting unknown pixels; then, they directly estimate the alpha mattes by comparing the similarity between the pixels to be predicted and the sample pixel pairs. Sampling-based methods can be further divided into two categories: (1) Local sampling-based methods, in which the samples are collected near the edge of the known regions in the image. Then, the alpha mattes are estimated according to the sample pixel pairs. This method has the advantage of having a smaller sampling range and faster processing speed. However, as the sampling space is confined to the edge of the known regions, this method may overlook the optimal sample pixel pair for the pixels to be predicted, leading to a lower accuracy in the predicted alpha mattes; and (2) Global sampling-based methods, which collect samples from all known regions and focus on the trade-off among the sample set size, coverage of the optimal pixel pair, and design of the pixel pair evaluation functions. Collecting samples from all known labeled regions in the images can effectively compensate for the shortcomings of a method using local sampling and improve the accuracy of alpha mattes. However, the processing speed decreases as the number of samples increases, and the sampling quality is unstable as the collected samples may also miss the optimal pixel pairs in complex situations.

3.2. Propagation-Based Methods

Propagation-based methods [3,50–53], also known as affinity-based methods, initially establish connections between adjacent pixels and then use a specific optimization strategy to propagate opacity information from known regions to unknown regions, leveraging this information to estimate the alpha mattes. Propagation-based methods can be divided into three categories: the one-level color smooth model, two-level alpha-color model, and three-level color mixture model. The one-level color smooth model calculates the alpha values of unknown pixels based on the assumption of color smoothness, which means that the foreground and background images of an arbitrary image are locally smooth and that each pixel in a small region of the image can be represented by a linear combination of adjacent pixels. The two-level alpha-color model estimates the alpha values based on the alpha-color assumption, which implies that alpha values can be generated by combining the corresponding color channels linearly. The three-level color mixture model is similar to the one-level color smooth model as it also relies on the color approximation of pixels in a small local region of images. However, in the three-level color mixture model, pixels that need to be predicted are represented as a linear combination of all pixels within a region. This suggests that each pixel is expressed as a mixture of several similar pixels.

The advantages of the methods that are based on propagation are as follows. (1) The alpha information is propagated from known regions to unknown regions, ensuring that the estimated alpha matte has local smoothness characteristics to avoid the problem of discontinuous alpha mattes caused by sampling methods. (2) The propagation of alpha information between unknown and known pixels can be modeled as an optimization problem based on quadratic objective functions, which can be solved efficiently using mature mathematical optimization tools, resulting in low computational cost. However, the working principle of propagation-based methods also leads to a considerable decrease in the performance of extracting the foregrounds of discontinuous objects. Moreover, the space complexity of propagation-based methods significantly increases as the image resolution increases as they need to store the correlations between pixels. This limits their application in high-resolution images.

3.3. Hybrid Methods

Hybrid methods [54–58] that incorporate both sampling and propagation techniques typically use a quadratic cost function that takes into account the smoothness of the alpha matte as well as the user input. The alpha matte is estimated by solving a sparse linear system of equations. There are two ways to implement a hybrid method: (1) the smoothness refinement method, which uses sampling-based strategies to initially estimate the alpha mattes and then utilizes propagation-based strategies to further refine the alpha mattes; and (2) the non-local affinity method, which is an extension of local affinity-based methods that searches for pixels with similar features to the target unknown pixel, establishes non-local affinity connections between them, and propagates alpha information from known regions to unknown regions.

Hybrid methods can effectively enhance the accuracy of alpha mattes by leveraging the strengths of both sampling-based and propagation-based approaches. However, this improved accuracy often comes at the cost of increased computational complexity, as these methods require performing both sampling and propagation operations. Therefore, a balance between accuracy and efficiency needs to be carefully considered when selecting an appropriate method for image matting.

4. Deep Learning Methods

Compared with traditional methods that rely on low-level and manual features to extract alpha mattes, deep learning methods leverage extensive datasets to train neural networks that can automatically learn image features and predict high-quality alpha mattes without requiring manual intervention. This approach has significantly improved alpha matting performance and has become increasingly popular recently. Deep learning-based image matting algorithms can be categorized into two categories: (1) algorithms improving the shortcomings of traditional algorithms through deep learning; and (2) end-to-end deep learning algorithms that generally achieve better performance than the former. Therefore, this paper focuses on recent innovative and breakthrough end-to-end deep learning algorithms, which can be categorized into two categories: trimap-based methods, which use a trimap as an additional input; and trimap-free methods, which do not require a trimap. The relevant methods mentioned in this article and their major innovations will be presented in Table 1.

Table 1. The major innovations and contributions of relevant image matting methods.

Method	Major Innovations and Contributions
DIM [9]	The first model to implement end-to-end neural networks for image matting.
AlphaGAN [12]	The first generative adversarial network in image matting.
VDRNet [17]	The introduction of a residual structure in the encoder helps to address the issue of performance degradation that arises with increasing network depth.
LSM [15]	The proposed method has improved the accuracy of alpha mattes by reducing unknowns through the calculation of foreground and background colors and by combining sampling methods with deep learning techniques.
MatteFormer [39]	The first method that introduced a transformer into image matting.
GCA [24]	The proposed model with an additional module named the guided contextual attention module effectively addresses the issue of poor matting performance in semi-transparent areas of images.
IndexNet [16]	The proposed approach has enhanced image details by learning index functions and leveraging them to guide the down-sampling and up-sampling processes.
SIM [25]	Expanding the traditional trimap to a semantic trimap that can provide various semantic information for the model.
DisIM [13]	Proposed a network that consists of one encoder and two decoders to identify the real mixed pixels in the unknown regions and estimate their alpha values simultaneously.

Table 1. *Cont.*

Method	Major Innovations and Contributions
CAMatting [14]	Proposed a two-encoder-two-decoder network to estimate the alpha matte and foreground image simultaneously.
HRMatting [26]	An approach specifically designed for high-resolution images, which involves running matting by cropping and stitching image patches. This approach can effectively process high-resolution images.
LFPNet [27]	Proposed a two-stream structure to learn long-range features outside the receptive field that can help distinguishing local foreground and background.
DAPM [8]	This is the pioneering end-to-end matting algorithm that can perform the task without the need for a trimap.
SHM [10]	Proposed the first two-stage CNN matting model, which uses a semantic segmentation model to generate a trimap and inputs it into the matting model to generate the alpha matte.
JMNet [22]	Introduction of a posture estimation module to provide global structural guidance and local focus of attention.
SalientIM [31]	Combined a salient object detection model to achieve a matting algorithm for the arbitrary salient foreground.
LFM [18]	Employed two decoder branches to make predictions on the foreground and background of an image, which are subsequently fed into the fusion module to generate the alpha matte.
GFM [36]	Designed a network consisting of a shared encoder and two decoders, which serve to learn semantic features and extract low-level structures to preserve details.
PP-Matting [35]	Similar to the idea of GFM [36], a guidance module is designed on this basis, which uses semantic features to provide guidance for detail prediction.
P3M [30]	The first algorithm designed for privacy-preserving portrait matting.
AlphaNet [20]	Developed an attention module that directs the up-sampling and down-sampling processes of the matting network, refining the alpha matte edges.
HATTMatting [19]	Combining appearance cues and pyramidal features through the integration of both spatial and channel attention mechanisms; uses hybrid loss functions to guide the network to improve foreground structure.
AIM [29]	Used attention to generate specific trimaps for different foregrounds, guiding the matting network to improve the details in unknown regions.
BGM [21]	Proposed a matting algorithm that uses a background image as an auxiliary input to generate the alpha matte instead of a trimap.
BGMV2 [28]	Implemented a real-time video matting algorithm using the background as an auxiliary input; Runs at 60 fps in HD on modern GPUs.
ModNet [33]	Proposed a CNN-based efficient and real-time human video matting method that does not require a trimap.
RVM [34]	Introduced ConvGRU module to utilize temporal information and achieved a robust real-time video matting algorithm.
AM [11]	Proposed a novel active learning approach that utilizes reinforcement learning to automatically determine the most informative labeling regions, while also obtaining a satisfactory alpha matte in user interaction.
SmartScribbles [23]	The model is able to effectively identify informative regions in the image and use information propagation methods to reduce the user scribbles. It can generate a high-quality alpha matte based on limited user scribbles.
UGIM [38]	An end-to-end CNN is designed in this method that supports arbitrary forms of trimap inputs, from completely unknown trimaps to carefully edited ones. The foreground and background regions of the trimmed map can be guided by user scribbles.
IIM [32]	This method generates heatmaps as prior information based on the user's clicks. The model predicts the foreground and background based on the click location, uses uncertainty estimation for fine-tuning adjustment, and then generates the alpha matte.
DIIM [37]	The model converts clicks into Euclidean distance maps and inputs them along with the origin image to the CNN for predicting the alpha matte.

4.1. Trimap-Based Methods

Trimaps can provide significant prior knowledge for deep neural networks, making them a popular choice as auxiliary inputs to reduce the difficulty of model training in many methods. This section provides a detailed introduction to representative trimap-based methods, which can be roughly categorized into three categories based on their network structure: a one-stream structure, one-stream structure with additional modules, and multi-stream structure.

4.1.1. One-Stream Structure

The one-stream structure was the earliest deep neural network structure used in image matting, and it remains a very popular structure to this day. Xu et al. [9] highlighted a limitation of the prior techniques, which solely rely on low-level features. The efficacy of such approaches might decline if the colors of the foreground and background are alike or if the texture of the image is complex. To overcome this challenge, the authors proposed a deep encoder-decoder neural network architecture that can automatically learn high-level features from input images and use them to predict the alpha matte. This groundbreaking study provided valuable insights for future research directions in image matting. Building upon the work of [9], Lutz et al. [12] introduced the use of generative adversarial networks (GANs) to tackle image matting challenges. The authors designed a generator-discriminator model to solve the matting task, where the generator is an encoder-decoder network; this innovative approach significantly improved the performance of image matting, demonstrating the vast potential of GANs in image matting. Expanding on the work of [9], Tang et al. [17] proposed several improvements. The authors developed an enhanced encoder-decoder network architecture to obtain more comprehensive foreground objects and introduced a residual structure in the encoder to address the performance degradation that arises with increased network depth. These advancements have significantly enhanced the accuracy and robustness of image matting.

Tang et al. [15] utilized a one-stream deep neural network to estimate the foreground and background of the input image before estimating the opacity values. The proposed method has conceptual similarities to sampling-based methods where color samples are selected for each pixel from the foreground and background; additionally, an initial alpha matte is obtained using Equation (1). However, the implementation of sample selection is different in this method as it trains a deep neural network specifically for sample selection. This approach utilizes the structure and texture information of the image by initially estimating foreground and background colors, thereby providing a more dependable estimation of the alpha matte.

In recent years, transformers have made significant progress in natural language processing and have shown a potential to replace CNNs in some computer vision tasks. Park et al. [39] were the first to introduce transformers into image matting and proposed a transformer-based one-stream matting architecture. The architecture employs both spatial tokens and prior tokens for self-attention, where prior tokens, which are global representations of each region in trimap, are utilized through prior-attention swin transformer modules to address the issue of an insufficient receptive field caused by the self-attention mechanism in local regions. Multiple outputs are initially generated in the alpha matte generation stage, which is selectively fused to gradually refine the unknown regions and generate a fine alpha matte. This research demonstrates the promising potential of transformers in image matting.

4.1.2. One-Stream Structure with Additional Modules

Although the one-stream structure is simple and has relatively low computational complexity, it often lacks sufficient auxiliary information, leading to poor performance in some complex scenarios. To overcome this limitation, some methods have proposed additional modules to augment the model with more contextual and structural information.

Li et al. [24] observed that previous methods often resulted in an unclear structure or texture of semi-transparent regions due to the local blurriness of transparent objects. To address this issue, they developed an encoder-decoder network with an additional module named guided contextual attention module. Specifically, their approach involves guiding alpha information from the image context to unknown pixels and using low-level feature information to propagate alpha features. Additionally, they proposed a guided contextual attention module that can distribute opacity information globally by relying on the learned low-level affinity information, leading to an improved performance in semi-transparent regions. These advancements offer a promising approach for improving the matting quality of transparent objects.

Lu et al. [16] discovered that index-guided decompression performs better than other up-sampling operators in restoring edge details during the decoding process of neural networks. The authors suggested a unification of current up-sampling operators with index functions, which can be treated as a function of the feature map that the network model can adaptively learn from the data. Specifically, they designed a one-stream model with a custom additional module, called IndexNet, which can dynamically predict and learn indexes based on the given inputs. The learned index function can then guide the down-sampling and up-sampling processes in the network to achieve better performance without requiring additional supervision.

Sun et al. [25] proposed that previous methods had not adequately considered the potential impact of semantic information of the foreground on the matting effect. Therefore, the authors proposed expanding the traditional trimap to a semantic trimap, which includes semantic information. They introduced a patch-based classifier to extract semantic information from the input image and merge it with the original trimap to create the semantic trimap. The semantic trimap and input image are fed into the u-structure for learning. Additionally, a multi-class discriminator and gradient constraint with weights sensitive to the content were introduced. The utilization of a multi-class discriminator regulates alpha mattes at a semantic level, while the inclusion of a gradient constraint with content-sensitive weights balances the various regularization losses. The utilization of a semantic trimap enhances the prior knowledge of the model and reduces the search space of the model when reasoning foreground objects.

4.1.3. Multi-Stream Structure

Because the one-encoder-one-decoder structure can only address a single problem, several matting methods have introduced a multi-stream structure to tackle multi-task problems. By utilizing these approaches, the network is able to acquire and extract a wider range of features from the input image, leading to enhanced flexibility and performance for various tasks.

According to Cai et al. [13], inferring the alpha matte directly from the trimap can lead to a decrease in algorithm performance due to difficulties in identifying the real mixed pixels in the unknown regions and estimating their alpha values simultaneously. In order to tackle this problem, the authors put forward a new method that decomposes the matting task into two subtasks and developed a one-encoder-two-decoder model to solve both problems simultaneously. This approach enables better utilization of semantic information and provides the model with additional structure awareness and a fault-tolerant capability.

Hou et al. [14] proposed a context-aware method for image matting in natural settings, which allows for the simultaneous estimation of the foreground and alpha matte. This method investigates the impact of both local regions and global contextual information on high-quality matting and designs a network that contains two encoders and two decoders and integrates global sampling and local propagation strategies to estimate the context-aware foreground and alpha matte in parallel. Through the fusion of regional visual features and global contextual information, the neural network provides sufficient feature information for estimating the alpha matte, effectively improving its accuracy.

Most matting methods have shown excellent performance on test datasets, but their practical performance can be less impressive due to the high resolution of images and due to hardware limitations. Therefore, Yu et al. [26] proposed a method specifically designed for high-resolution images, which operates by cropping and stitching image patches while explicitly modeling contextual information across them, addressing the issues of context dependency and consistency between different image patches. Specifically, the proposed method employs a dual encoder structure to extract features from both the image block and trimap block. The resulting feature blocks are then processed using a custom cross-patch context module, which captures long-term contextual dependency information across image blocks, making stitching boundaries smoother after restoring all image blocks. This approach was the first to address the challenges of high-resolution image matting.

Liu et al. [27] discovered that the effective receptive field of ordinary convolutional neural networks (CNNs) is often too small to connect pixels in unknown areas with pixels in known areas, which can significantly affect the performance of the model. To address this issue, the authors designed a two-stream structure model based on the crop and patch approach; one branch called the matching module was used to process cropped image blocks, and the other branch called the propagating module was utilized to process context image blocks centered around the cropped image blocks. This processing strategy allows for the learning of long-range features outside the receptive field, which helps in distinguishing the local foreground and background and in solving the issues mentioned above, ultimately improving the overall performance of the model.

4.1.4. Summary of Trimap-Based Methods

The introduction of trimaps provides clear foreground and background information for the deep neural network such that the model can focus on the calculation of unknown regions, which optimizes the training effect of the model and allows the model to infer high-quality alpha mattes with fine edges. However, the difficulty of generating a trimap and the demand for matting-related knowledge limit the application of trimap-based methods.

4.2. Trimap-Free Methods

The underconstrained nature of the image matting problem necessitates that a trimap be used in most methods as an auxiliary input to add constraints that are difficult to apply in the real world. Trimaps are also difficult to deal with in large-scale data or in tasks with timelines (e.g., videos, live streaming). Therefore, in recent years, many image matting methods without auxiliary input have been proposed; these methods are called trimap-free methods. Trimap-free methods are more applicable to the real world. This section introduces recent superior and breakthrough trimap-free methods.

4.2.1. Methods for Automatically Generating Trimaps

Due to the complexity involved in creating trimaps, it is unfriendly to users that do not have knowledge of matting. As a result, many researchers have begun exploring methods for automatically generating trimaps to avoid user interaction. This approach makes image matting techniques more accessible and user-friendly.

Shen et al. [8] proposed a method for automatic portrait matting. The authors regarded the trimap prediction as a semantic segmentation task with three classes (i.e., background, foreground, and transition regions), using FCN [59] for prediction and facial landmark points to align the mask. Additionally, they proposed a matting layer that can perform forward and backward propagation, achieving an end-to-end automatic portrait matting system. Shen et al. [8] only used a CNN model in the trimap generation stage, whereas Chen et al. [10] utilized CNN separately for both the trimap generation and matting tasks. The authors also introduced a fusion module to preserve the semantic information of the foreground and background, allowing the results generated by two distinct networks (semantic segmentation and image matting) to blend with each other. Wu et al. [22] proposed a new portrait matting method that differs from the sequential approach used

by [8,10]. The authors used Resnet50 [60] to serve as the backbone, which was coupled with the ASPP [61] module for the extraction of multi-scale features. These features were subsequently fed into the human pose network, trimap network, and matting network individually. With the semantic information generated by the human pose network assisting the trimap and matting networks, the authors achieved better results. The resulting alpha matte was more refined compared with previous methods. As trimap-free methods heavily rely on expensive alpha matte annotations, most of the existing methods are targeted at portrait matting. Therefore, Deora et al. [31] proposed a framework for salient object matting that utilizes the U2Net [62] network to generate a trimap for any salient object present in an image. This trimap is then passed to the matting model for further refinement, effectively solving the problem of previous trimap-free methods being only capable of single-object matting. As a result, this approach expands the application scenarios of trimap-free models.

4.2.2. Dual-Decoder Structure

To achieve a trimap-free method without using a semantic segmentation network to generate trimaps, researchers have designed a one-stage dual-stream network structure. This parallel decoder network structure makes the model simpler and easier to train compared with the two-stage network [8,10,22,31] while achieving equally precise visual effects. Zhang et al. [18] extracted semantic features from the image using a dual decoder and predicted the foreground and background separately. As the semantic segmentation results are coarse, the authors designed a network that blends the foreground and background to produce a refined alpha matte without generating a trimap, which achieves excellent results. In contrast to [18], which predicted the foreground and background, Li et al. [36] proposed a division of the matting task into two simultaneous subtasks: a high-level semantic segmentation alongside the preservation of low-level structural features. The two decoders respectively learn semantic features to distinguish the foreground and background and extract low-level structures to preserve details. The two decoders work together to complete the matting task. Chen et al. [35] share a similar idea to Li et al. [36] through the use of two decoders for the semantic learning and detail extraction of the transition region. The authors added a guidance module to provide guidance for detail prediction using semantic features, achieving interaction between detail and semantic features. This enables the network to generate an alpha matte with excellent detail effects in the transition region.

4.2.3. Attention Method

Attention mechanisms have demonstrated significant effectiveness and power as a tool in the field of computer vision, which can help models better focus on key information in images. In the image matting domain, many CNN models have adopted attention mechanisms to obtain more precise alpha mattes. We will now introduce the relevant research works of other researchers in detail.

Sharma et al. [20] proposed an attention module that projects the entirety of the feature map's channels onto a singular attention map. This map is adaptively learned from the training data and guides the up-sampling and down-sampling processes of the matting network for better boundary detail enhancement. Qiao et al. [19] proposed a hierarchical attention-based matting network that utilizes spatial attention to capture image textures and appearance clues; the method also uses channel attention to extract semantic features, similarly to some dual decoder networks [35,36]. Li et al. [29] classified foreground types in images and employed SE attention [63] to learn the semantic features of different foreground types, enabling the attention to guide the matting network to improve detail in the transition regions of different foreground types.

4.2.4. User-Friendly Input

The process of creating a trimap requires a significant amount of time and effort. When there is no trimap available, trimap-free methods often extract all salient objects in an image,

making it challenging to extract specific foregrounds based on user requirements. Therefore, designing a more user-friendly auxiliary input or interactive segmentation methods would meet user demands and also have high research value.

Sengupta et al. [21] proposed a portrait matting method using a background image as an auxiliary input, which replaces the trimap. By designing a network architecture, this method simultaneously achieves the output alpha matte and foreground image, further improving the efficiency of the matting process. Then, Lin et al. [28] modified the network structure based on [21] to achieve real-time video matting and proposed a new idea of using the foreground residual as the prediction target. Experimental results indicate that the utilization of the foreground residual can substantially enhance the convergence rate of the network and effectively address color overflow problems.

The difficulty of obtaining a background image is much lower than that of creating a trimap; however, usually only the background image can be conveniently obtained in the real world, and when processing videos, this method's effectiveness significantly decreases when the background changes. In addition to using the background to replace the trimap, researchers have designed models where users provide constraint information through interaction. Yang et al. [11] designed a recursive reinforcement learning model that can detect the marked regions in the image to obtain a good alpha matte and can provide feedback for detecting the next suggested marked region based on the foreground and background of user scribbles. Through iteration, the model obtains a fine alpha matte. On the other hand, Qiao et al. [23] divided the image into blocks and automatically selected important regions for generating the alpha matte based on the image content. Users scribbled the foreground, background, and transition regions, and the CNN expanded the scribbled information to the entire image to iteratively generate an accurate trimap. These models combine user interaction with traditional matting algorithms to generate the final alpha matte. Unlike the approach of combining interactive information with traditional matting algorithms [11,23], Fang et al. [38] implemented a CNN architecture that supports arbitrary trimap inputs. In the absence of a trimap input, the model assumes a completely unknown trimap as the input and generates the corresponding foreground and background. The model can also be guided by user scribble markings to correct the generated alpha matte. Wei et al. [32] and Ding [37] et al. used user clicks as auxiliary inputs. Wei et al. [32] marked the foreground and background of the image by utilizing user clicks. The corresponding heat map was generated for the click position and was inputted into the model to generate the alpha matte. In addition to using clicks to mark the foreground and background, Ding et al. [37] also added markings for the unknown area, which would generate the corresponding Euclidean distance map. The model gradually propagated the marking information to adjacent areas through a specific strategy to continuously correct the alpha matte generated by the model.

4.2.5. Real-Time Video Matting

Real-time scenarios, such as video conferences and live streaming, commonly utilize image matting techniques. Therefore, designing a lightweight and efficient video matting model framework that operates at high resolutions without requiring auxiliary inputs is an important research direction for scholars.

Lin et al. [28] utilized the background to assist image matting, which increases the user threshold and limits the application scenarios. Therefore, Ke et al. [33] proposed MODNet, a lightweight network that employs branch networks for edge refinement and achieves real-time video matting without auxiliary inputs. MODNet also addresses the issue of flickering in the video matting process, resulting in output results that are more in line with human perception. After proposing [28], Lin et al. went on to propose a real-time video matting model [34] that operates independently without requiring auxiliary inputs. This model utilizes ConvGRU [64] to leverage temporal information in videos and uses DGF [65] modules to enhance image resolution, thereby improving the robustness and

matting quality of the model. The RVM [34] model highlights the importance of temporal consistency and quality in video matting models.

4.2.6. Summary of Trimap-Free Methods

The trimap-free methods utilize the powerful feature extraction ability of deep neural networks to generate alpha mattes without the need for a trimap. Researchers have continuously improved the quality of trimap generation using semantic information, attention mechanisms, or user interaction. As these models do not require the production of a trimap, the user's expertise requirement is greatly reduced, and the application scenarios for image matting models are becoming more widespread. However, it is important to note that without the guidance of a trimap, these algorithms may be less robust in certain scenarios. For instance, in images with multiple objects or moving objects, the models may struggle to accurately separate the foreground from the background. This limitation arises because the absence of a trimap makes it challenging for the algorithm to differentiate between different objects or track their movements effectively.

5. Evaluation Indicators

The evaluation indicators of image matting are employed to assess the quality of alpha mattes, which reflects the quality of the corresponding image matting algorithm. Image matting evaluation techniques can be divided into two groups: subjective evaluation and objective evaluation. Due to the subjectivity of subjective evaluation, it is generally not used. Therefore, this section aims to introduce objective evaluation indicators and briefly outline their usage methods.

Objective evaluation indicators measure the disparity between algorithm-generated alpha mattes and ground truth alpha mattes to assess the quality of image matting, and they can be classified into single-point and multi-point evaluation methods depending on the range considered for the opacity value difference.

5.1. Single-Point Evaluation

For single-point evaluations, the discrepancy between the ground truth alpha mattes and alpha mattes estimated by the algorithms is calculated for each pixel using the mean square error (MSE) and sum of absolute difference (SAD). To mathematically represent the calculation of the discrepancy between the estimated alpha matte and ground truth alpha matte for pixel i, the MSE is expressed as

$$MSE = \frac{1}{N}\sum_{i}^{N}(\hat{\alpha}_i - \widetilde{\alpha}_i)^2. \tag{2}$$

The smaller the MSE, the better the model fits the dataset. The MSE has strong robustness to changes in brightness and contrast, but it gives higher weight to errors with larger pixel values, which may lead to less accurate evaluation results.

The sum of absolute value errors is denoted by

$$SAD = \sum_{i}^{N}|\hat{\alpha}_i - \widetilde{\alpha}_i|. \tag{3}$$

where N represents the number of unknown pixels in the unknown region of the trimap, while $\hat{\alpha}_i$ and $\widetilde{\alpha}_i$ denote the alpha values of pixel i in the estimated and ground truth alpha mattes, respectively.

The smaller the value of the SAD, the higher the degree of fitting between the model and the dataset. The SAD is sensitive to changes in brightness and contrast and therefore is not suitable for evaluating image quality in complex scenes.

Single-point evaluation indicators such as MSE and SAD are simple to use and have low computational costs. However, they may not accurately capture the subjective

perception of image quality by the human eye, which limits their ability to reflect the true performance of algorithms in practical applications.

5.2. Multi-Point Evaluation

A single-point evaluation is a straightforward method to calculate but fails to reflect the relationship between the opacity differences of multiple pixels. Therefore, in specific cases, a multi-point evaluation is necessary. Multi-point evaluations take into account the differences between the ground truth alpha mattes and algorithm-generated alpha mattes from the perspective of a series of pixels, including gradient error and connectivity error [40], among others. The gradient error reflects the discrepancy in the transparency alteration between the ground truth alpha matte and the algorithm-generated alpha matte and is calculated as

$$Grad = \sum_i (\bigtriangledown \hat{\alpha}_i - \bigtriangledown \widetilde{\alpha}_i)^q. \tag{4}$$

where $\bigtriangledown \hat{\alpha}_i$ and $\bigtriangledown \widetilde{\alpha}_i$ denote the normalized gradient magnitude of the estimated alpha matte and ground truth alpha matte at pixel i, respectively; they are calculated using the Gaussian first derivative. q is a custom parameter.

The algorithmic matting performance is deemed superior when $Grad$ attains a smaller value. The efficacious evaluation of algorithms in terms of details (e.g., hair, fur, etc.) can be achieved via the gradient error.

The connectivity error represents the difference in connectivity between the ground truth alpha matte and the estimated alpha matte, and it is mathematically defined as

$$Conn = \sum_i (\varphi(\hat{\alpha}_i, \Omega) - \varphi(\widetilde{\alpha}_i, \Omega))^p. \tag{5}$$

where Ω represents the largest connected region in both the ground truth alpha matte and the estimated alpha matte, which is the foreground. p is a custom parameter. φ is a function that characterizes the extent of connectivity between pixel i and Ω, which is calculated as

$$\varphi(\hat{\alpha}_i, \Omega) = 1 - (\lambda_i \cdot \delta(d_i \geq \theta) \cdot d_i). \tag{6}$$

Pixels are fully disconnected when $\varphi = 0$ and fully connected when $\varphi = 1$. λ_i corresponds to the mean distance between pixel i and the nearest connected pixel to Ω: $\lambda_i = \frac{1}{|K|} \sum_{k \in K} dist_k(i)$, K is the set of discrete values between l_i and α_i. The $disk_k$ function computes the normalized Euclidean distance between pixel i and the nearest pixel connected to Ω when using a threshold of k. In the equation $d_i = \hat{\alpha}_i - l_i$, l_i refers to the maximum binarization threshold that is needed to ensure four-connectivity between pixel i and Ω; δ denotes the Dirac δ function, which is used to ignore d_i being less than θ. θ is a custom parameter where pixels are considered fully connected when $d_i < \theta$, resulting in a more flexible error calculation. The smaller the $Conn$, the smoother the surface changes of the foreground object, the clearer the boundaries, and the better the matting effect.

However, as neither the gradient error nor the connectivity error alone can comprehensively assess the quality of the alpha matte, a combination of both indicators is often employed to evaluate the performance of a model.

5.3. Utilizing of Evaluation Indicators

The deep learning-based algorithms of image matting predict the alpha matte of unlabeled images using training models to learn image features from labeled data. During the training process, evaluation indicators are typically used to supervise and optimize the performance of the model. These evaluation indicators can reflect various differences between the predicted alpha mattes of the model and the ground truth alpha mattes. The training objective aims to minimize the values of these indicators, narrow the disparities between the algorithm-generated alpha mattes and the ground truth alpha mattes, and optimize the performance of the model.

Specifically, a loss function such as MSE can be employed to quantify the discrepancy between the alpha mattes generated by the model and the ground truth alpha mattes; then, optimization algorithms such as gradient descent can be used to minimize the value of this loss function as much as possible. During the training process, we input the training data into the model, calculate the predicted alpha matte of the model, and input the estimated alpha matte and corresponding ground truth alpha matte into the loss function to calculate the difference. Subsequently, an optimization algorithm is utilized to update the model parameters. This process continues to iterate until the difference reaches the optimum or converges to an acceptable value.

6. Applications of Image Matting

As mentioned earlier, image matting is a widely used image processing technique that has numerous applications in various scenarios. It involves image editing, medical imaging, cloud detection, and other related fields. In this section, we discuss the application scenarios where image matting plays a significant role as well as the fundamental principles behind their applications.

Image Editing: Image matting is a versatile tool for fine-tuning and manipulating images to achieve various desired effects in image editing. Foreground separation is a key application of image matting, which enables image composition and finds use in diverse fields such as advertising design and video production [66,67]. Additionally, image matting is useful in image inpainting [68], which is a technique for repairing damaged or missing parts of an image. Beyond these, image matting can also be utilized for background replacement and object movement; with the former, the background of an image can be replaced with another image or color, while the latter allows for the movement of the foreground object to another position. These applications can be used to achieve a range of visual effects for the image.

Medical Imaging: In the field of medical imaging, image matting technology has found applications in the segmentation and analysis of medical images. For example, Wang et al. [69] introduced alpha mattes of image matting into medical imaging, which better described lesion details and addressed the blurring problem associated with binary masks. Moreover, image matting can also be utilized for medical image segmentation and disease detection [70,71]; the former can be used to separate various structures or tissues (such as lungs and bones) in medical images, assisting in analysis and diagnosis, while the latter can detect diseased tissues in medical images and help to locate lesion locations.

Cloud Detection: The utilization of image matting technology can be extended to the domain of cloud detection [72–74]. Traditional techniques encounter challenges when attempting to accurately segment clouds in remote sensing images due to their uneven brightness and irregular shapes. However, image matting offers a more superior performance by separating the foreground and background of clouds, resulting in improved accuracy and efficiency in cloud detection.

Aside from the mentioned applications, image matting is widely applicable in various other domains. For instance, in game development, this technology can facilitate the free movement and operation of game characters by separating them from the background. In intelligent transportation systems, image matting allows for the precise detection and tracking of vehicles or pedestrians by separating them from the background. Additionally, image matting is extensively utilized in multi-modality, automatic driving, and 3D applications.

7. Challenges and Future Directions

7.1. Constructing Large-Scale Datasets

Datasets play a critical role in deep learning methods, and high-quality datasets such as COCO [41] and Imagenet [75] have made a significant impact on related research fields. However, due to the laborious process of annotating alpha mattes, most publicly available matting datasets are constructed using synthetic techniques. These involve synthesizing

different images by combining foregrounds with different backgrounds to effectively expand the dataset. However, the limited variety of foreground types and the introduction of artefacts in the synthesized images can result in poor learning outcomes, leading to overfitting and a reduced generalization ability in neural networks.

Furthermore, due to the expensive production costs, current hand-annotated image matting datasets focus on a single category, typically humans or animals. Natural multi-class image matting datasets are therefore rare. Some recent studies, such as GFM, have proposed methods to improve image synthesis that address issues such as resolution and noise. These studies have also constructed unique background datasets to enhance the quality of synthetic datasets.

We believe that proposing new data augmentation and synthesis strategies to overcome the limitations of synthetic datasets in image matting is crucial. Additionally, exploring how advanced image generation algorithms can assist in constructing matting datasets is an important area for future research. In summary, building large-scale high-quality image matting datasets that include diverse foregrounds and backgrounds as well as precise alpha mattes is necessary to advance the field of image matting.

7.2. Multi-Class Matting

Currently, image matting is a binary classification problem of foreground and background. Exploring how to add more class information on this basis to achieve multi-class classification without auxiliary input, similar to semantic segmentation, is a new research direction that is worth exploring. This will greatly enhance the application scenarios and flexibility of image matting because it can freely choose the foreground category that needs to be extracted. For example, in an image of a little girl holding a dog, current models can only extract the girl and dog as the foreground together rather than separately extracting her pet dog, which is very inconvenient. With multi-class matting, desired foreground objects can be extracted through semantic information. Research on multi-class matting will provide more value for the industrial application of image matting.

7.3. Robust Trimap-Free Matting

The current trimap-free methods combined with the powerful feature extraction capability of CNN and increasing number of open-source datasets can already generate fine alpha mattes without the need for a trimap. However, they still have limitations, and their performance is not ideal in some special scenarios. We randomly experimented with some open-source trimap-free matting models [29,30,34,35], and the results are shown in Figure 6. When processing moving foreground objects, the edges of the objects in motion may blend and overlap with the background, creating blurry areas that greatly increase the difficulty of image processing. The model fails to match them correctly. In addition, current models also struggle to perfectly extract small foreground objects with different scales when they appear in images. This is because most matting datasets contain images with large foreground objects that dominate the image. The model tends to overlook smaller foreground objects within an image when there are foreground objects of varying scales present. It is currently a major challenge for trimap-free methods to extract foreground objects with completely different scales and to deal with motion blur.

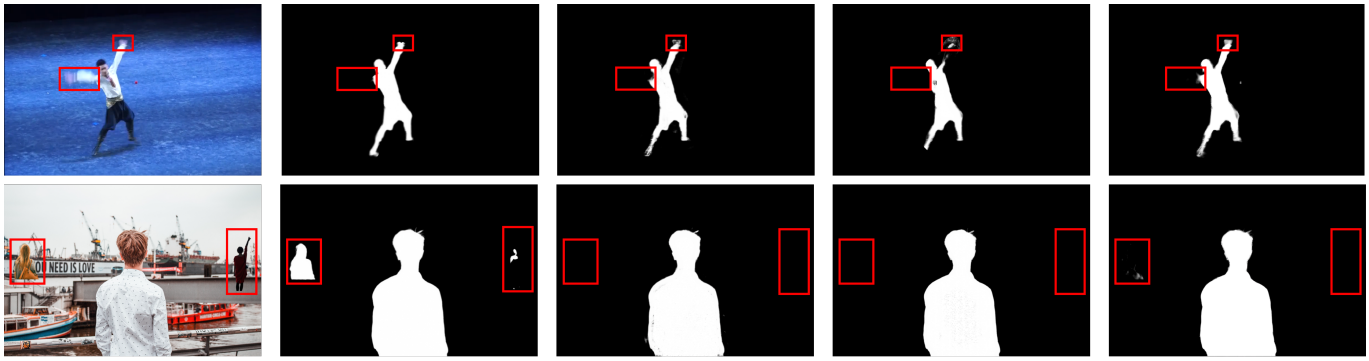

Figure 6. The first row displays the model's output alpha mattes for a motion-blurred image, while the second row displays the model's output alpha mattes for different scales in the foreground image, where (**a**) represents the output of the PPM [35] model, (**b**) represents the output of the AIM [29] model, (**c**) represents the output of the P3M [30] model, and (**d**) represents the output of the RVM [34] model. As shown in the figure, the methods perform poorly in capturing the motion-blurred objects indicated by the red rectangles in the first row. Additionally, when objects with different scales, as indicated by the red rectangles in the second row, appear in the image, these methods generally fail to extract them accurately and tend to only capture the most prominent objects.

8. Conclusions

In conclusion, image matting is a fundamental problem in computer vision and has garnered significant attention from the research community. Over the years, numerous approaches have been proposed to solve this problem, including traditional methods based on manual and low-level features as well as more advanced deep learning-based techniques. In this paper, we have provided an extensive survey of image matting methods based on deep learning. We first introduced well-known datasets and traditional methods. Subsequently, we concentrated on summarizing notable image matting methods based on deep learning and analyzing their innovations. Additionally, we discussed typical evaluation indicators for image matting quality, along with several applications of the technique. Finally, we explored the primary challenges and future directions of image matting. Despite significant progress, image matting remains a vibrant research topic, and we believe that future work in this area will undoubtedly drive advancements in computer vision and related fields.

Author Contributions: Conceptualization, L.H. and X.L.; methodology, L.H., X.L. and B.T.; software, L.H.; validation, L.H., X.L. and B.T.; formal analysis, L.H., X.L. and B.T.; resources, B.T.; writing—original draft preparation, L.H. and X.L.; writing—review and editing, B.T., X.W. and J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Guangxi Science and Technology Major Project under Grant No. AA22068057 and the Youth Science Fund Project of Guangxi Natural Science Foundation under Grant No. 2021GXNSFBA220039.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors thank everyone that provided suggestions and assistance for this paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Smith, A.R.; Blinn, J.F. Blue screen matting. In Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, New Orleans, LA, USA, 4–9 August 1996; pp. 259–268.
2. Mishima, Y. Soft Edge Chroma-Key Generation Based Upon Hexoctahedral Color Space. US Patent 5,355,174, 11 October 1994.
3. Sun, J.; Jia, J.; Tang, C.K.; Shum, H.Y. Poisson matting. In *ACM SIGGRAPH 2004 Papers*; ACM: New York, NY, USA, 2004; pp. 315–321.
4. Li, X.; Li, J.; Lu, H. A survey on natural image matting with closed-form solutions. *IEEE Access* **2019**, *7*, 136658–136675. [CrossRef]
5. Boda, J.; Pandya, D. A survey on image matting techniques. In Proceedings of the 2018 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 3–5 April 2018; pp. 765–770.
6. Yao, G.L. A survey on pre-processing in image matting. *J. Comput. Sci. Technol.* **2017**, *32*, 122–138. [CrossRef]
7. Wang, J.; Cohen, M.F. Image and video matting: A survey. *Found. Trends Comput.* **2008**, *3*, 97–175. [CrossRef]
8. Shen, X.; Tao, X.; Gao, H.; Zhou, C.; Jia, J. Deep automatic portrait matting. In Proceedings of the Computer Vision ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 92–107.
9. Xu, N.; Price, B.; Cohen, S.; Huang, T. Deep image matting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2970–2979.
10. Chen, Q.; Ge, T.; Xu, Y.; Zhang, Z.; Yang, X.; Gai, K. Semantic human matting. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Republic of Korea, 22–26 October 2018; pp. 618–626.
11. Yang, X.; Xu, K.; Chen, S.; He, S.; Yin, B.C.; Lau, R. Active matting. *NeurIPS* **2018**, *31*; pp. 4595–4605.
12. Lutz, S.; Amplianitis, K.; Smolic, A. Alphagan: Generative adversarial networks for natural image matting. *arXiv* **2018**, arXiv:1807.10088.
13. Cai, S.; Zhang, X.; Fan, H.; Huang, H.; Liu, J.; Liu, J.; Liu, J.; Wang, J.; Sun, J. Disentangled image matting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8819–8828.
14. Hou, Q.; Liu, F. Context-aware image matting for simultaneous foreground and alpha estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4130–4139.
15. Tang, J.; Aksoy, Y.; Oztireli, C.; Gross, M.; Aydin, T.O. Learning-based sampling for natural image matting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3055–3063.
16. Lu, H.; Dai, Y.; Shen, C.; Xu, S. Indices matter: Learning to index for deep image matting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3266–3275.
17. Tang, H.; Huang, Y.; Fan, Y.; Zeng, X. Very deep residual network for image matting. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 4255–4259.
18. Zhang, Y.; Gong, L.; Fan, L.; Ren, P.; Huang, Q.; Bao, H.; Xu, W. A late fusion cnn for digital matting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 7469–7478.
19. Qiao, Y.; Liu, Y.; Yang, X.; Zhou, D.; Xu, M.; Zhang, Q.; Wei, X. Attention-guided hierarchical structure aggregation for image matting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13676–13685.
20. Sharma, R.; Deora, R.; Vishvakarma, A. AlphaNet: An attention guided deep network for automatic image matting. In Proceedings of the 2020 International Conference on Omni-Layer Intelligent Systems (COINS), Barcelona, Spain, 31 August–2 September 2020; pp. 1–8.
21. Sengupta, S.; Jayaram, V.; Curless, B.; Seitz, S.M.; Kemelmacher-Shlizerman, I. Background matting: The world is your green screen. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2291–2300.
22. Wu, X.; Fang, X.N.; Chen, T.; Zhang, F.L. JMNet: A joint matting network for automatic human matting. *Comput. Vis. Media* **2020**, *6*, 215–224. [CrossRef]
23. Yang, X.; Qiao, Y.; Chen, S.; He, S.; Yin, B.; Zhang, Q.; Wei, X.; Lau, R.W.H. Smart scribbles for image matting. *ACM Trans. Multimed. Comput.* **2020**, *16*, 121. [CrossRef]
24. Li, Y.; Lu, H. Natural image matting via guided contextual attention. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11450–11457.
25. Sun, Y.; Tang, C.K.; Tai, Y.W. Semantic image matting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 11120–11129.
26. Yu, H.; Xu, N.; Huang, Z.; Zhou, Y.; Shi, H. High-resolution deep image matting. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 3217–3224.
27. Liu, Q.; Xie, H.; Zhang, S.; Zhong, B.; Ji, R. Long-range feature propagating for natural image matting. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual, 20–24 October 2021; pp. 526–534.
28. Lin, S.; Ryabtsev, A.; Sengupta, S.; Curless, B.L.; Seitz, S.M.; Kemelmacher-Shlizerman, I. Real-time high-resolution background matting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 8762–8771.
29. Li, J.; Zhang, J.; Tao, D. Deep automatic natural image matting. *arXiv* **2021**, arXiv:2107.07235.

30. Li, J.; Ma, S.; Zhang, J.; Tao, D. Privacy-preserving portrait matting. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual, 20–24 October 2021; pp. 3501–3509.
31. Deora, R.; Sharma, R.; Raj, D.S.S. Salient image matting. *arXiv* **2021**, arXiv:2103.12337.
32. Wei, T.; Chen, D.; Zhou, W.; Liao, J.; Zhao, H.; Zhang, W.; Yu, N. Improved image matting via real-time user clicks and uncertainty estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 15374–15383.
33. Ke, Z.; Sun, J.; Li, K.; Yan, Q.; Lau, R.W.H. Modnet: Real-time trimap-free portrait matting via objective decomposition. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 1140–1147. [CrossRef]
34. Lin, S.; Yang, L.; Saleemi, I.; Sengupta, S. Robust high-resolution video matting with temporal guidance. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 238–247.
35. Chen, G.; Liu, Y.; Wang, J.; Peng, J.; Hao, Y.; Chu, L.; Tang, S.; Wu, Z.; Chen, Z.; Yu, Z.; et al. Pp-matting: High-accuracy natural image matting. *arXiv* **2022**, arXiv:2204.09433.
36. Li, J.; Zhang, J.; Maybank, S.J.; Tao, D. Bridging composite and real: Towards end-to-end deep image matting. *Int. J. Comput. Vis.* **2022**, *130*, 246–266. [CrossRef]
37. Ding, H.; Zhang, H.; Liu, C.; Jiang, X. Deep interactive image matting with feature propagation. *IEEE Trans. Image Process* **2022**, *31*, 2421–2432. [CrossRef] [PubMed]
38. Fang, X.; Zhang, S.H.; Chen, T.; Wu, X.; Shamir, A.; Hu, S.M. User-Guided Deep Human Image Matting Using Arbitrary Trimaps. *IEEE Trans. Image Process* **2022**, *31*, 2040–2052. [CrossRef]
39. Park, G.; Son, S.; Yoo, J.; Kim, S.; Kwak, N. Matteformer: Transformer-based image matting via prior-tokens. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 11696–11706.
40. Rhemann, C.; Rother, C.; Wang, J.; Gelautz, M.; Kohli, P.; Rott, P. A perceptually motivated online benchmark for image matting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 1826–1833.
41. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
42. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
43. Adobe Inc. *Adobe Photoshop*, Version: CC 2019. Available online: https://www.adobe.com/products/photoshop.html (accesssed on 6 March 2019).
44. Chuang, Y.Y.; Curless, B.; Salesin, D.H.; Szeliski, R. A bayesian approach to digital matting. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, Kauai, HI, USA, 8–14 December 2001; Volume 2, p. 2.
45. Wang, J.; Cohen, M.F. Optimized color sampling for robust matting. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
46. He, K.; Rhemann, C.; Rother, C.; Tang, X.; Sun, J. A global sampling method for alpha matting. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011; pp. 2049–2056.
47. Gastal, E.S.L.; Oliveira, M.M. Shared sampling for real-time alpha matting. In *Computer Graphics Forum*; Blackwell Publishing Ltd.: Oxford, UK, 2010; Volume 29, pp. 575–584.
48. Shahrian, E.; Rajan, D. Weighted color and texture sample selection for image matting. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 718–725.
49. Shahrian, E.; Rajan, D.; Price, B.; Cohen, S. Improving image matting using comprehensive sampling sets. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 636–643.
50. Grady, L. Random walks for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 1768–1783. [CrossRef]
51. Levin, A.; Lischinski, D.; Weiss, Y. A closed-form solution to natural image matting. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *30*, 228–242. [CrossRef]
52. Zheng, Y.; Kambhamettu, C. Learning based digital matting. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 889–896.
53. Levin, A.; Rav-Acha, A.; Lischinski, D. Spectral matting. *IEEE Trans. Pattern Anal.* **2008**, *30*, 1699–1712. [CrossRef]
54. Muja, M.; Lowe, D. Fast approximate nearest neighbors with automatic algorithm configuration. *VISAPP* **2009**, *2*, 2.
55. Chen, X.; Zou, D.; Zhou, Z.; Zhao, Q.; Tan, P. Image matting with local and nonlocal smooth priors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 1902–1907.
56. Chen, Q.; Li, D.; Tang, C.K. KNN matting. *IEEE Trans. Pattern Anal.* **2013**, *35*, 2175–2188. [CrossRef]
57. Lee, P.; Wu, Y. Nonlocal matting. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011; pp. 2193–2200.
58. Aksoy, Y.; Aydin, T.O.; Pollefeys, M. Designing effective inter-pixel information flow for natural image matting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 29–37.
59. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.

60. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
61. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
62. Qin, X.; Zhang, Z.; Huang, C.; Dehghan, M.; Zaiane, O.R.; Jagersand, M. U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recogn.* **2020**, *106*, 107404. [CrossRef]
63. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
64. Siam, M.; Valipour, S.; Jagersand, M.; Ray, N. Convolutional gated recurrent networks for video segmentation. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 3090–3094.
65. Wu, H.; Zheng, S.; Zhang, J.; Huang, K. Fast end-to-end trainable guided filter. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1838–1847.
66. Niu, L.; Cong, W.; Liu, L.; Hong, Y.; Zhang, B.; Liang, J.; Zhang, L. Making images real again: A comprehensive survey on deep image composition. *arXiv* **2021**, arXiv:2106.14490.
67. Zhang, H.; Zhang, J.; Perazzi, F.; Lin, Z.; Patel, V.M. Deep image compositing. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 5–9 January 2021; pp. 365–374.
68. Zhao, Y.; Price, B.; Cohen, S.; Gurari, D. Guided image inpainting: Replacing an image region by pulling content from another image. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, 7–11 January 2019; pp. 1514–1523.
69. Wang, L.; Ye, X.; Ju, L.; He, W.; Zhang, D.; Wang, X.; Huang, Y.; Feng, W.; Song, K.; Ge, Z. Medical matting: Medical image segmentation with uncertainty from the matting perspective. *Comput. Biol. Med.* **2023**, *158*, 106714. [CrossRef]
70. Li, X.; Yang, T.; Hu, Y.; Xu, M.; Zhang, W.; Li, F. Automatic tongue image matting for remote medical diagnosis. In Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Kansas City, MO, USA, 13–16 November 2017; pp. 561–564.
71. Cheng, J.; Zhao, M.; Lin, M.; Chiu, B. AWM: Adaptive Weight Matting for medical image segmentation. In *Medical Imaging 2017: Image Processing*; SPIE: Bellingham, WA, USA, 2017; Volume 10133, pp. 769–774.
72. Li, W.; Zou, Z.; Shi, Z. Deep matting for cloud detection in remote sensing images. *IEEE Trans. Geosci. Remote* **2020**, *58*, 8490–8502. [CrossRef]
73. Ma, D.; Wu, R.; Xiao, D.; Sui, B. Cloud Removal from Satellite Images Using a Deep Learning Model with the Cloud-Matting Method. *Remote Sens.* **2023**, *15*, 904. [CrossRef]
74. Liu, J.; Chen, X.; Chen, Q.; Zheng, Q.; Fu, H.; Qian, J. Matting-based automatic and accurate cloud detection for multisource satellite images. *J. Appl. Remote Sens.* **2020**, *14*, 026519. [CrossRef]
75. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.