A Study of Meta-Learning Methods on the Problem of Video Matting

by

Negin Tabaraki

A thesis submitted to the
School of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements for the degree of

Masters of Science in **Computer Science**

Faculty of Science
University of Ontario Institute of Technology (Ontario Tech University)
Oshawa, Ontario, Canada
April 2024

THESIS EXAMINATION INFORMATION

Submitted by: **Negin Tabaraki**

Master of Science in Computer Science

Thesis Title: A Study of Meta-Learning Methods on the Problem of Video Matting

An oral defense of this thesis took place on April 11^{th}, 2024 in front of the following examining committee:

Examining Committee:

Chair of Examining Committee	Dr. Mehran Ebrahimi
Research Supervisor	Dr. Faisal Qureshi
Research Supervisor	Dr. Ken Pu
Examining Committee Member	Dr. Ali Neshati
Thesis Examiner	Dr. Steven Livingstone

The above committee determined that the thesis is acceptable in form and content and that a satisfactory knowledge of the field covered by the thesis was demonstrated by the candidate during an oral examination. A signed copy of the Certificate of Approval is available from the School of Graduate and Postdoctoral Studies.

Abstract

Applying image matting techniques directly to video matting presents challenges, primarily due to the complex temporal dynamics inherent in video data. In this work, we studied two Meta Learning approaches—Boosting with Adapters (BwA) and Boosting using Ensemble (BuE)—to tackle the task of video matting using pre-trained image matting models. BwA refines (image matting) alpha mattes by fine tuning pre-trained segmentation models, which we refer to as adapters, using video frames. BuE, additionally, combines multiple fine-tuned adapters using a convolutional neural network. We introduced a meta-learning architecture that incorporates both adapters and ensemble boosting through an iterative process of expert selection and fine tuning. Based on our evaluation on benchmarks based on a standard video matting dataset (VideoMatte240K), we confirm that the proposed scheme improves the performance of image matting models on the task of video matting. In addition, the proposed approach also improves the performance of VMFormer (c. 2022), a recent video matting method.

Keywords: video matting; alpha matte enhancement; adaptive segmentation models; meta-learning; ensemble;

Author's Declaration

I hereby declare that this thesis consists of original work of which I have authored. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners. I authorize the Ontario Tech University to lend this thesis to other institutions or individuals for the purpose of scholarly research. I further authorize the Ontario Tech University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research. I understand that my thesis will be made electronically available to the public.

Negin Tabaraki

Statement of Contributions

I hereby certify that I am the sole author of this thesis and I have used standard referencing practices to acknowledge ideas, research techniques, or other materials that belong to others. Furthermore, I hereby certify that I am the sole source of the creative works and/or inventive knowledge described in this thesis.

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Dr. Ken Pu and Dr. Faisal Qureshi, for their guidance, encouragement, and support throughout my research journey. their invaluable insights, constructive feedback, and unwavering dedication have been instrumental in shaping this thesis.

I would also like to extend my appreciation to my lab colleagues for their invaluable support and collaboration. Their valuable input, discussions, and feedback have greatly enriched my research experience and enabled me to broaden my horizons.

Finally, I would like to thank my family for their support, and encouragement. Their constant encouragement and belief in me have been a constant source of motivation and inspiration, and I am forever grateful for their unwavering support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In the world of digital media, there is a widespread and growing need for advanced visual content in different industries. This demand encompasses areas such as film production, Virtual and Augmented Reality (VR and AR), advertising, and online education [1]. All these fields require effective methods to produce realistic and engaging visual content. Video matting, a set of methods and techniques, responds to this need by allowing the separation of subjects from backgrounds in video frames. This process is essential for adding special effects, merging real and virtual footage, and enhancing overall visual quality. The role of video matting is central in not only improving the digital content but also in ensuring its interactive qualities. As the demand for digital media continues to rise, video matting come out as a key solution, addressing the challenges of modern visual content creation.

1.1 Motivation

The journey from image segmentation to video matting has been marked by significant advancements, each building on the foundation laid by earlier methods. Initially, image segmentation models, with their complex structures and extensive training on large datasets, paved the way for developments in image matting. These image mat-

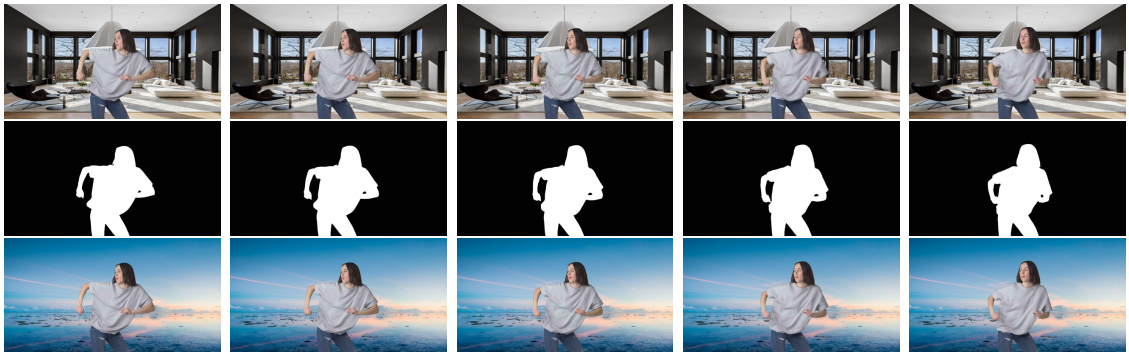


Figure 1.1: An example of video matting. Top row: Consecutive original frames. Middle row: Corresponding alpha mattes. Bottom row: Extracted frames with new backgrounds.

ting techniques, which focused on accurately separating foreground elements from static backgrounds, demonstrated the potential of applying sophisticated neural network architectures to solve complex visual tasks.

The availability of strong, pre-trained image matting models, trained on vast and diverse datasets, has provided a rich foundation for further innovation. These models have been instrumental in achieving high accuracy in image matting, setting a precedent for quality that is challenging to replicate in dynamic environments such as video sequences.

As the field progressed, the focus shifted from still images to videos, introducing the need for techniques that could handle the dynamic nature of video sequences. Researchers experimented with various architectures, including Long Short-Term Memory networks (LSTMs) and transformers [2, 3], to tackle the challenges of video matting, such as maintaining consistency across frames and managing changing scenes. Unlike image matting, video matting models are much more complex and resource-intensive to train from scratch, given their need to understand and process temporal dynamics alongside spatial information.

In our research, recognizing the challenges of training video matting systems and the existing capabilities of image matting models, we aimed to bridge this gap through meta-learning. Drawing inspiration from meta-learning, which enables models to quickly adapt to new tasks using limited data, we explored its application in adapting pre-

trained image matting models for video matting. Our goal was to develop a method that could leverage the existing strengths of image matting models and enhance them for video content, achieving high-quality matting results with less reliance on extensive computational resources and large datasets. This approach seeks to make video matting more accessible and efficient, harnessing the power of meta-learning to adapt robust image matting solutions to the more complex and dynamic domain of video matting.

1.2 Thesis Outline

In this thesis, Chapter 2 lays the groundwork by providing a background on the main concepts central to our study. It delves into the preliminaries necessary for understanding the subsequent chapters. Chapter 3 reviews both traditional and deep learning methods in segmentation, image, and video matting, offering a comprehensive overview of existing methodologies and developments in these areas. Chapter 4 is dedicated to our approach and methodology, where we detail the specific strategies and techniques employed to address the challenges of video matting. In Chapter 5, we present our experiments and the results obtained, showcasing the practical application and effectiveness of our methods. Finally, Chapter 6 concludes the thesis, summarizing our findings and discussing the limitations and potential avenues for future work in this field.

1.2.1 Contributions

Our thesis makes the following contributions to the field of video matting:

- **Adaptation of Image Matting to Video Contexts:** We extend image matting techniques to video matting by applying transfer learning. This method addresses the dynamic complexities of video, enhancing the adaptability of matting techniques to the temporal information found in video sequences.

- **Boosting with Adapters (BwA):** Our research introduces a novel boosting method that refines alpha mattes through the use of advanced segmentation models. This methodology significantly improves the accuracy and realism of video matting results.

- **Boosting using Ensemble (BuE):** We have developed an ensemble method that combines multiple fine-tuned adapters to enhance the overall quality of video matting. This approach boosts the robustness and effectiveness of matting systems across varied video content.

- **Multi-stage training** In our approach, we designed a two-stage process to enhance video matting. Initially, we fine-tuned adapters and selected the most effective ones based on their performance. Subsequently, we froze these adapters' weights and trained an ensemble layer to intelligently combine their outputs, significantly improving the overall quality of the matting results.

Chapter 2

Preliminaries

In this chapter, we introduce the key concepts essential to our research. This chapter serves as a groundwork, discussing topics such as image segmentation, image matting, video matting, and meta-learning techniques. Image segmentation and matting provide the basis for understanding how objects are delineated and extracted from images, a fundamental task that feeds into the more complex task of video matting. This extension into video matting is crucial, as it incorporates the challenges of motion and temporal continuity. Additionally, meta-learning techniques provide us with strategies for enhancing learning accuracy, essential for the tasks involved in video matting. This chapter establishes a solid theoretical base, preparing for the detailed research and analysis ahead in the thesis.

2.1 Image Segmentation

Image segmentation is a task in computer vision, where the goal is to partition an image into multiple segments, each corresponding to different objects or regions [4].

As we discussed in previous sections, a color image is represented as a tensor $\mathbf{I} \in \mathbb{R}^{3\times m\times n}$. In the context of image segmentation:

- The image is assumed to be composed of multiple objects $\{O_1, O_2, \ldots, O_n\}$.
- Each object O_i is a collection of pixels, forming a subset of the image.
- Formally, an object O_i can be represented as $O_i \in \mathcal{P}(m \times n)$, where $\mathcal{P}$ denotes the power set of all pixel coordinates in image $\mathbf{I}$.

To model an object O as a tensor, we use a learning model $\mathbf{M} \in \{0,1\}^{m \times n}$ where:

$$(i,j) \in O \iff \mathbf{M}[i,j] = 1.$$

This model $\mathbf{M}$ represents a mask for the object, with 1 indicating the presence of the object at pixel location (i,j).

2.1.1 Segmentation Task

The task of image segmentation involves determining the mask for one or more objects present in an image [4]. If there are Z objects, then a training sample can be represented as a pair $(\mathbf{I}, \mathbf{M})$, where I is the input image and $\mathbf{M}$ is the output multi-object mask in $\{0,1\}^{Z \times m \times n}$.

A neural network solution for image segmentation might be structured as follows:

- **Input**: $\mathbf{I} \in \mathbb{R}^{3 \times m \times n}$, representing the image.
- **Output**: $\mathbf{p} \in [0,1]^{Z \times m \times n}$, representing the probabilities of object presence.
- For each pixel location (i,j) and object k, $\mathbf{p}[k,i,j] \in [0,1]$ denotes the probability that pixel (i,j) belongs to object O_k.

This problem can be solved with a multi-layer neural network. The specific architecture and configuration of the network can be different based on the requirements of the segmentation task. In the related work section, we will discuss several existing methods

for image segmentation using deep neural networks which have been shown to be effective in segmentation tasks.

2.1.2 Evaluating Image Segmentation Model

The effectiveness of an image segmentation model can be quantified using a loss function. For image segmentation tasks, particularly those involving binary classification of pixels, the Binary Cross Entropy (BCE) loss [5] is commonly used.

Binary Cross Entropy Loss

The binary cross entropy loss for a multi-object segmentation task is computed over each object and each pixel in the image. Given a training sample with the ground truth mask $\mathbf{M} \in \{0,1\}^{F\times m\times n}$ and the predicted probability map $\mathbf{p} \in [0,1]^{F\times m\times n}$, the BCE loss L is calculated as follows:

$$L = -\sum_{k=1}^{F}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(\mathbf{M}[k,i,j]\log(\mathbf{p}[k,i,j]) + (1-M[k,i,j])\log(1-\mathbf{p}[k,i,j])\right).$$

Here, for each object k and pixel location (i,j), $\mathbf{M}[k,i,j]$ is the ground truth label (0 or 1), and $\mathbf{M}[k,i,j]$ is the predicted probability of belonging to the object O_k. The binary cross entropy loss is effective in segmentation tasks as it directly measures the discrepancy between the predicted probabilities and the actual labels at the pixel level. It encourages the model to predict probabilities that are close to the ground truth binary labels, thereby improving the accuracy of the segmentation.

2.2 Image Matting

Image matting is a task in computer vision, focusing on the precise extraction of foreground elements from the background of an image. This section outlines the mathematical formulation and the neural network approach for image matting. We will also explore traditional methods and neural network approaches in greater detail in the related work section.

2.2.1 Mathematical Formulation of Image Matting

In image processing, an image denoted as $\mathbf{I}$ is represented as a three-dimensional tensor within the space $\mathbb{R}^{3\times m\times n}$. A key task in image matting is the computation of an alpha matte, symbolized as α. This alpha matte is essentially a two-dimensional matrix where each element corresponds to a pixel in the image, with values ranging from 0 to 1. Here, the value $\alpha(i, j)$ indicates the transparency degree of the pixel located at position (i, j), where 0 represents complete transparency (entirely background) and 1 signifies complete opacity (entirely foreground).

The matting equation, as described in [6], articulates the relationship between the image, the foreground, the background, and the alpha matte through the formula:

$$I(i, j) = \alpha(i, j) \cdot F(i, j) + (1 - \alpha(i, j)) \cdot B(i, j),$$

where $\mathbf{F}(i, j)$ and $\mathbf{B}(i, j)$ represent the foreground and background colors at the pixel (i, j), respectively. This equation posits that the color of each pixel in the image is a blend of the foreground and background colors, modulated by the alpha values. A practical challenge in image matting is accurately determining the alpha matte α when only the image $\mathbf{I}$ is known, while the foreground $\mathbf{F}$ and background $\mathbf{B}$ are unspecified.

The development of image matting began with traditional techniques that relied heavily on user input. Techniques such as manual trimap creation and color sampling were

the mainstays, requiring significant manual effort and struggling with complex textures. With advances in computational capabilities, semi-automated tools came out, reducing the reliance on manual input. Techniques like Bayesian matting [7] and graph cut-based algorithms offered more image understanding, improving the handling of edges and semi-transparent areas but still facing challenges with intricate details and varied lighting.

The use of deep learning significantly advanced the field of image matting. Neural networks, particularly convolutional neural networks (CNNs), began to be used for direct alpha matte prediction, excelling in handling complex details and diverse conditions.

Today, image matting is continuously developing, with research focusing on more efficient architectures, handling of edge cases, and reduction of high-quality data dependency. Developments like the integration of depth maps and generative adversarial networks (GANs) are at the forefront of current advancements.

In the related work section, we will present specific works and advancements in image matting, examining key milestones and state-of-the-art techniques in greater detail.

2.3 Video Matting

Video matting extends the principles of image matting to the dynamic and temporal aspects of video content. Unlike static images, a video comprises a sequence of frames, each with its unique challenges in matting.

2.3.1 Mathematical Framework for Video Matting

- **Video Representation**: A video V is a series of images over time, represented as $V = \{\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_t\}$.

- **Sequential Alpha Mattes**: For each frame $\mathbf{I}_t$ in $\mathbb{R}^{3\times m\times n}$, a corresponding alpha matte α_t, $\alpha_t \in [0,1]^{m\times n}$, is computed. This results in a set of alpha mattes $\{\alpha_1, \alpha_2, \ldots, \alpha_t\}$ for the video.

One of the main challenges in video matting is ensuring the temporal consistency of the alpha mattes across the video sequence [8]. Temporal consistency implies that the alpha mattes of adjacent frames, α_t, should not exhibit abrupt. This is crucial for achieving a seamless and realistic foreground extraction in the video playback. Mathematically, this can be expressed as ensuring the consistency of α_t with α_{t-1} and α_{t+1}, particularly in regions of the frame experiencing motion or other changes. Achieving this consistency often involves integrating temporal information into the matting algorithm. Techniques such as motion vectors or optical flow are used to adjust α_t based on the observed changes from frame $\mathbf{I}_{t-1}$ to $\mathbf{I}_t$. This approach helps in maintaining the natural flow and realism of the extracted foreground over time.

In the related work section, we will delve deeper into specific works and research that have contributed to the field of video matting, highlighting key methodologies and breakthroughs in both traditional and deep learning approaches.

2.4 Meta Learning Methods

Meta-learning [9] offers a powerful framework for enhancing machine learning models by aiming to improve their overall learning accuracy. This concept is particularly promising for video matting because it allows the integration of multiple weaker learners to create a stronger, more robust system through fine-tuning. By utilizing meta-learning techniques, we can leverage existing model checkpoints, which can then be fine-tuned to adapt effectively to specific video matting tasks. This ability to adapt quickly and efficiently makes meta-learning an ideal approach for handling the dynamic and complex nature of video sequences.

In this research, we explore two specific meta-learning strategies: using residual networks and boosting (an ensemble method). These approaches are chosen for their potential to enhance learning from limited data examples and their ability to generalize across

various video matting scenarios. By integrating these meta-learning techniques, we aim to address the challenges of video matting and push the boundaries of what current systems can achieve, both in terms of performance and efficiency.

2.4.1 Meta-Learning using Residual Networks

Residual learning [6] focuses on training layers to model a residual function, which refers to the difference between the desired output and the current output of the network. Suppose we have a target function $f(x)$ and a weak learner $M_\theta(x)$ that is currently unable to accurately approximate $f(x)$. This inadequacy is quantified by a significant loss, $\text{loss}(f(x), M(x))$. To enhance $M_\theta(x)$, we introduce a residual layer R_θ as follows:

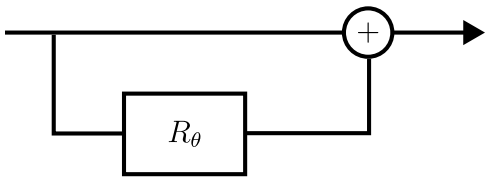


Figure 2.1: Diagram of a Residual Block

The architecture of R_θ can be designed based on the specific requirements of the task. A simple version of $R_\theta(x)$ could be a linear function, such as $R_\theta(x) = Ax + B$, with parameters $\theta = (A, B)$. In this setup, if both A and B are zero, $R_\theta(x)$ effectively becomes zero. However, more complex designs can incorporate convolutional layers, pooling layers, and non-linear activation functions like ReLU. These additions enable the residual layer to learn more complex modifications to the initial learner's output.

We can compose the main learner and the residual layer as follows:

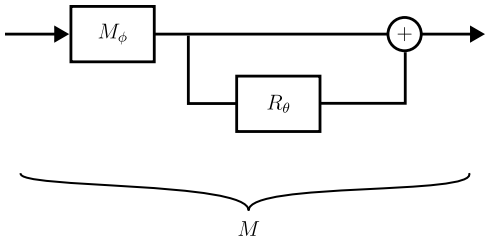


Figure 2.2: Composition of a Main Learner and a Residual Layer in a Meta Learning Model.

We define θ^* as the model parameter for which $R_{\theta^*}(x) = 0$. This configuration ensures that the residual layer does not alter the output of $M_\phi(x)$ when it is not necessary. The main learner M_ϕ and the residual layer R_θ can then be composed to form an enhanced model. If $\theta = \theta^*$, then the model M is equivalent to M_0, ensuring that the enhanced model is at least as accurate as the original weak learner M_ϕ.

The integration of the residual layer allows for the improvement of the weak learner M_θ through an iterative learning process. By applying gradient descent to $\nabla_\theta \text{Loss}(f(x), M(x))$, the model learns the optimal parameters of the residual layer. This process involves adjusting θ to minimize the loss, thereby learning the residuals needed to correct the predictions of $M_\theta(x)$ and bring them closer to $f(x)$.

The use of residual layers in neural networks offers several advantages. Primarily, it allows for the enhancement of existing models without the need to increase their complexity significantly. This method is particularly beneficial in scenarios where adding more layers or increasing the depth of the network is not feasible due to computational constraints. Furthermore, residual learning can help alleviate the problem of vanishing gradients in deep networks, making it easier to train deeper architectures.

2.4.2 Meta-Learning using Boosting

Another method in meta-learning, particularly relevant in complex classification scenarios, is boosting with the ensemble method. This approach addresses the limitations inherent in using a single binary classifier, especially when dealing with non-linear or

high-dimensional input spaces. Linear classifiers often struggle in these scenarios due to the non-linear nature of separation boundaries.

Boosting algorithms, such as AdaBoost [10] and XGBoost [11], begin with a basic premise: starting with a weak learner, an algorithm that performs slightly better than random chance. The strategy then involves incrementally adding more weak learners to this ensemble. Each new learner is tasked with correcting the errors made by its predecessors. This iterative process enables the ensemble to progressively improve its ability to classify complex data accurately.

The ensemble [12], often conceptualized as a decision tree, navigates through different choices based on the input data. The ensemble layer within this tree decides which weak models, such as M_1 or M_2, are best suited for the given input, leading to one of several possible outputs. This process is akin to navigating through a decision tree, where each branch represents a decision made by a particular weak model.

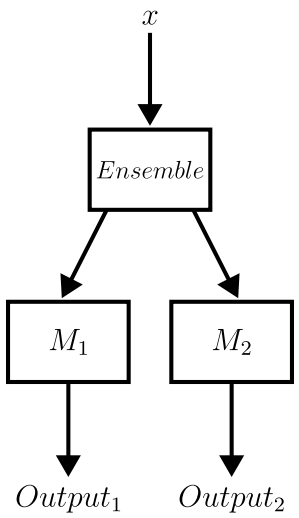


Figure 2.3: Schematic Representation of an Ensemble Decision Process.

Furthermore, this decision tree can be modeled as a neural network where the ensemble layer combines inputs, denoted as y_1 and y_2, based on the initial input x, to arrive at the final output y. This representation illustrates how the ensemble method integrates multiple weak learners to formulate a more accurate classification decision.

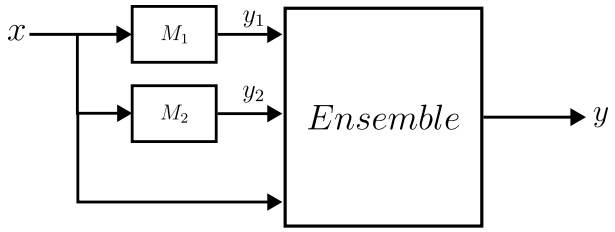


Figure 2.4: Illustration of an Ensemble Neural Network.

Central to the boosting process is the decision function, defined as select-best-model(x). This function is responsible for selecting the most appropriate weak model for a given input x. The final output is then derived from the selected model $M_i(x)$. Importantly, this function is not predefined but is learned during the training phase. It evolves as a black-box function, optimized based on the training data, to effectively determine the best model for any given input.

In essence, boosting with the ensemble method exemplifies the power of meta-learning in complex classification tasks. By systematically combining the capabilities of multiple weak learners, it creates a robust and adaptive classification system capable of handling intricate, non-linear, and high-dimensional data, overcoming the constraints of individual linear classifiers.

Chapter 3

Related Work

In the realm of computer vision, the tasks of image segmentation and matting play pivotal roles in understanding and processing visual information. Image segmentation, the process of partitioning an image into multiple segments, lays the groundwork for more advanced tasks such as image matting. Image matting extends this concept further by precisely extracting foreground elements from the background, usually in still images. This technique is crucial for isolating subjects in an image, allowing for detailed editing and manipulation.

The principles of image matting are then extrapolated to the more dynamic and complex medium of videos, leading to the task of video matting. Video matting involves separating foreground elements from backgrounds in video sequences, requiring a high degree of consistency across sequential frames – a challenge that is uniquely presented in the video format as opposed to still images.

In recent years, the field has undergone a significant transformation, marked by a shift from traditional methods to more sophisticated techniques, especially with the advent of deep learning. Deep learning approaches, grounded in the principles of machine learning, have shown remarkable capabilities in both image segmentation and matting. They offer enhanced results, particularly when dealing with complex scenarios in image and video

formats.

This section aims to provide a comprehensive overview of the progression in image segmentation, image matting, and video matting techniques. It will trace the evolution from early conventional methods to the latest advancements in deep learning, highlighting how each stage has contributed to the field's development and the challenges they address in computer vision.

3.1 Image Segmentation

3.1.1 Traditional Methods

Before the introduction of deep learning, traditional image segmentation methods were widely used. These methods can be classified into several categories:

1. **Thresholding Methods:** This category represents the simplest form of image segmentation. It involves dividing an image into foreground and background segments by setting a threshold value. Depending on this threshold, pixels are classified as either belonging to the foreground or the background [13]. There are two types of thresholding: global, which uses a single threshold value for the entire image, and adaptive, which applies different threshold values across different image areas [14].

2. **Edge-Based Methods:** These methods focus on detecting discontinuities in an image's brightness or color, which are indicative of edges [15]. Techniques such as the Canny, Sobel, and Prewitt edge detectors are commonly used for identifying these edges, which then aid in segmenting the image into various regions [16].

3. **Region-Based Methods:** In this approach, segmentation is based on identifying regions within the image that share similar characteristics [17]. The key techniques here are region growing, which starts from a seed point and expands by adding neighboring pixels with similar properties, and region splitting and merging, which

involves dividing the image into regions and then combining or separating them based on predefined criteria [18].

4. **Clustering Methods:** These methods involve grouping pixels into clusters based on certain properties like color, intensity, or texture [19]. Common clustering algorithms used for this purpose include K-means, Fuzzy C-means, and Mean-Shift.

5. **Histogram-Based Methods:** Utilizing the image's histogram, these methods segment an image by identifying significant peaks, valleys, and slopes within the histogram, which correspond to different regions in the image [20]. This approach is particularly effective in images with distinct intensity levels.

While traditional methods were effective for simpler applications, they had limitations in handling complex images, especially where the variation in object appearance, shape, and size was significant. These methods often required manual tuning of parameters and were not robust against noise and variations in lighting conditions. Also, traditional methods generally lacked the ability to learn from data, making them less adaptable and flexible compared to deep learning-based approaches.

In the following, we will explore how the advent of deep learning revolutionized image segmentation, offering more powerful, adaptable, and robust methods.

3.1.2 Deep Learning Methods

The limitations of traditional image segmentation methods have led to the adoption of deep learning techniques, notably convolutional neural networks (CNNs), which excel in handling complex real-world images [21]. Deep learning models automatically learn to identify patterns and structures from large datasets, eliminating the need for manual feature selection or parameter tuning.

These models excel in semantic segmentation, classifying each pixel into predefined categories, essential in applications like medical imaging or autonomous driving. They

are also robust against common image processing challenges, such as lighting and scale variations, due to their training on diverse datasets. The success of deep learning in image segmentation is largely due to the availability of large annotated datasets and advanced computational resources, marking a significant advancement in the field.

In the following sections, we'll explore the deep learning models and techniques in image segmentation.

DeepLabv3

DeepLabv3 is a key development in the use of deep learning for semantic image segmentation, well-known for its atrous convolution technique [22]. This approach allows the model to understand different scales in an image without losing detail. An important feature in DeepLabv3 is the use of atrous convolution, which broadens the filter's view, making it better at segmenting objects. It also includes an atrous spatial pyramid pooling (ASPP) module. This module examines the image at various levels, helping the model deal with objects that are different in size. DeepLabv3's design, which combines atrous convolution and ASPP, sets it apart in tasks that need detailed segmentation across scales. Its strong performance in tests shows it's effective for complex segmentation needs.

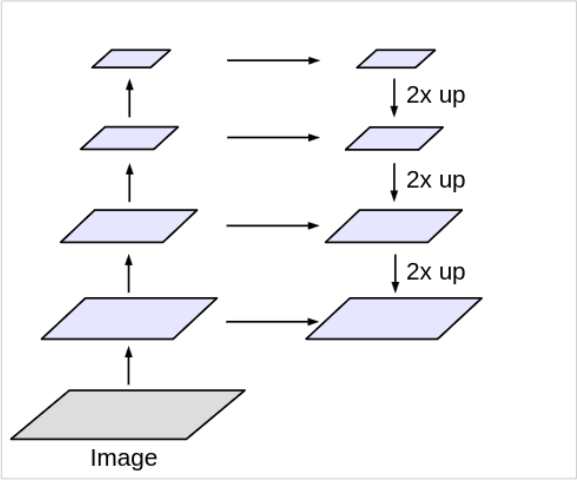


Figure 3.1: DeepLabV3's Architecture. This figure is taken from [23].

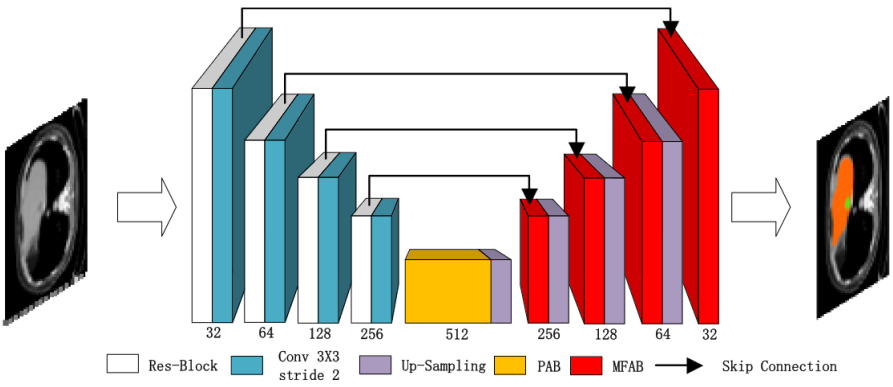


Figure 3.2: Multi-scale Attention Net's Architecture. This figure is taken from [24].

Multi-scale Attention Net

MANet, standing for Multi-scale Attention Net, brings a new approach to image segmentation by using attention mechanisms [24]. Its key feature is focusing on different levels of detail in an image, which helps in segmenting both small and large objects accurately. This multi-scale attention is particularly helpful when the image has objects of various sizes or when it's hard to tell objects apart from the background. MANet is versatile and performs well in complex situations, making it useful in areas like medical imaging for detailed analysis and in autonomous vehicle technology for recognizing different objects in a scene.

Pyramid Attention Network

The Pyramid Attention Network (PAN) is another development in using deep learning for image segmentation. Its main feature, the pyramid attention mechanism, blends both local and global context to improve the way features are represented, leading to better segmentation performance [25]. PAN's pyramid attention module works with the feature map at various scales, capturing different levels of context. This allows the model to effectively concentrate on the important features, no matter their size or where they are in the image. By merging local details with an overall view, PAN can accurately segment complex images, even those with diverse object sizes and detailed backgrounds.

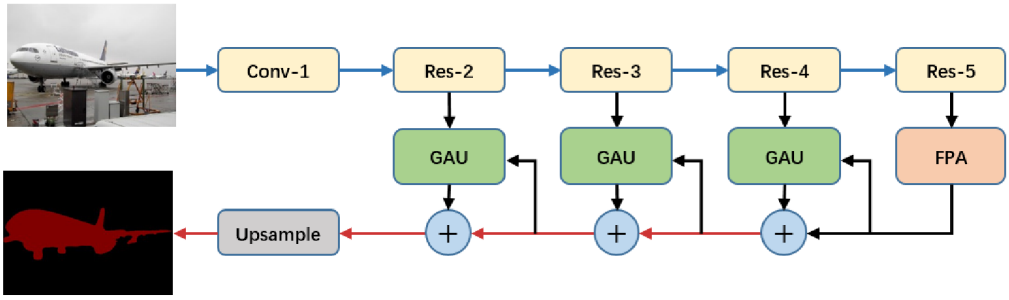


Figure 3.3: Pyramid Attention Network's Architecture. This figure is taken from [25].

LinkNet

LinkNet is another deep learning model known for its fast and efficient image segmentation, especially in real-time processing [26]. Its structure is based on an encoder-decoder architecture, designed to be quick without losing accuracy in segmentation. The main feature of LinkNet is its link connections. These connections effectively merge features from both the encoder and decoder parts of the network. This setup allows for fast movement of information and gradients, which is crucial for applications that need real-time processing. The encoder part of LinkNet is responsible for pulling out features from the input image, and the decoder then uses these features to build the segmentation map.

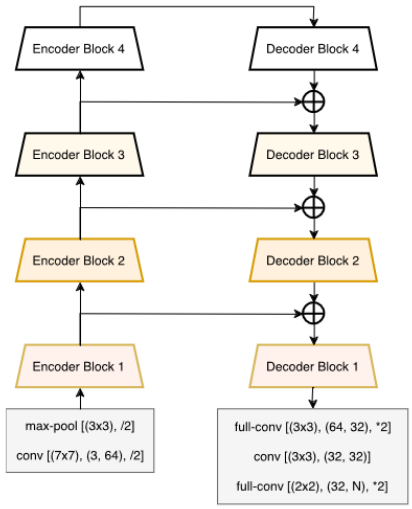


Figure 3.4: LinkNet's Architecture. This figure is taken from [26].

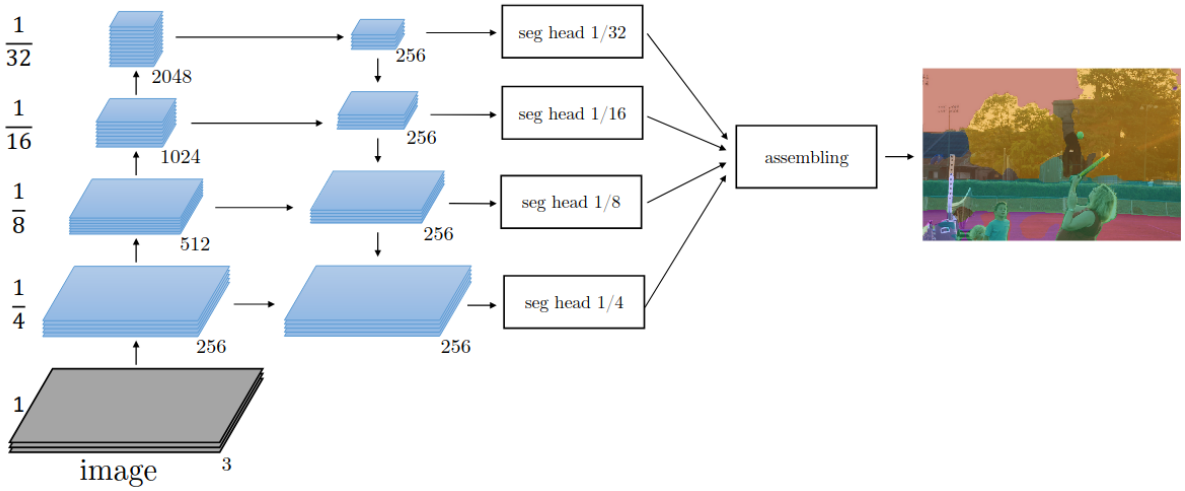


Figure 3.5: Feature Pyramid Network's Architecture. This figure is taken from [27].

Feature Pyramid Networks

Feature Pyramid Networks (FPN) are also important in image segmentation, especially known for dealing with objects of different sizes [27]. FPN creates a feature pyramid from a single image, which helps in detecting and segmenting objects no matter their scale. What makes FPN innovative is its unique top-down architecture with side connections. This design mixes features that are low in resolution but high in meaning with those that are high in resolution but lower in meaning. This combination allows the network to create detailed feature maps at all scales, enhancing its ability to segment images with objects of various sizes.

UNet

UNet is a widely recognized model in image segmentation, particularly valued for its use in medical image analysis [28]. UNet's design is unique for its symmetric structure, which helps in accurately identifying specific locations and understanding the context in images, both crucial for precise segmentation. The model consists of two main parts: a contracting path that captures the overall context, and a symmetric expanding path that

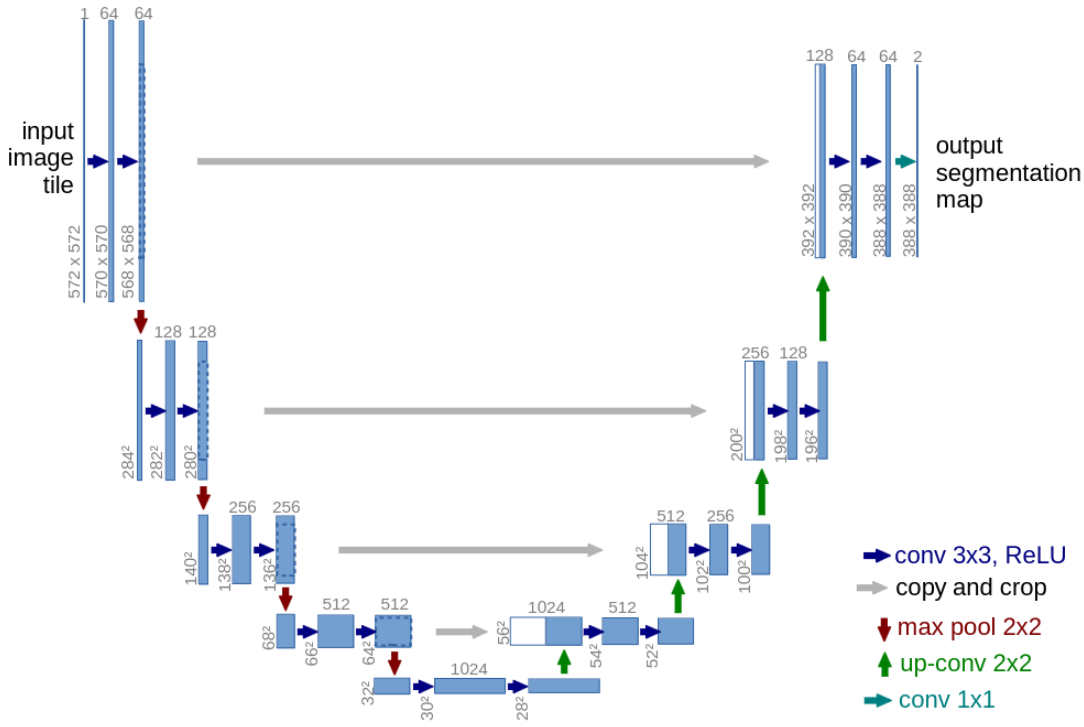


Figure 3.6: UNet's Architecture. This figure is taken from [28].

focuses on precise localization. What enhances UNet's design are the skip connections that link these two paths. These connections allow detailed context information to be carried over to the higher resolution layers. This feature is key for combining simpler, low-level features with more complex, high-level ones, leading to more detailed and accurate segmentation.

3.2 Image Matting

3.2.1 Traditional Methods

At the early stages of image matting research, techniques based on samples were leading the field. Berman et al [29] introduced a novel method using a cursor to select dominant colors in images. This technique was particularly effective, as it produced several candidate mattes for each pixel, skillfully addressing the complexities in semi-transparent

areas of images. In a distinct approach, Chuang et al. [7] adopted a Bayesian method. Their research focused on estimating the opacity of each pixel, integrating a probabilistic approach into sample-based methods.

As research in image matting progressed, there was a gradual shift towards methods based on propagation. A notable contribution by Sun et al. [30] redefined the matting problem as a task of solving Poisson equations. This shift brought notable improvements, particularly in dealing with complex scenes. At the same time, Levin et al. [31] explored a deterministic path. Their work involved formulating a cost function to analytically solve the matting problem. Their method was based on how the foreground and background colors in an image are related.

3.2.2 Deep Learning-Based Methods

The progress in both sample-based and propagation-based methods created a foundation for deep learning to enter the field of image matting. The goal was to increase accuracy and efficiency. However, before 2017, deep learning was not much used for image matting due to hardware limitations and not having enough data to train the deep learning models effectively.

Trimap-Based methods

The significant turning point in image matting came with the study by Xu et al. [6] Their research showed the potential of deep learning in image matting and tackled the existing limitations. They employed an encoder-decoder matting network combined with a smaller refinement network, achieving unparalleled results on both the alphamatting.com benchmark [32] and their Composition-1K benchmark [6], which later became a key standard for evaluating image matting.

A key contribution of their work was the creation of the Adobe Deep Matting dataset, which initially had 431 images along with their corresponding ground truth alpha mattes.

By using these alpha mattes to extract foregrounds and composite them onto different backgrounds, Xu et al. [6] created thousands of new composite images. They further enhanced the dataset by introducing random augmentations to the image-trimap pairs, such as random crops, flipping, and trimap dilation, to promote better algorithm generalization. The new algorithm presented by Xu et al. marked a significant advancement in image matting technology, overcoming the shortcomings of earlier algorithms that relied heavily on low-level features and lacked high-level context.

Following this foundational work, other researchers built upon and researched on deep learning-based image matting methods. Cai et al. [33] broke down the matting pipeline into two sub-tasks: trimap adaptation and alpha matte estimation, achieving state-of-the-art results. They first inferred global structural semantics on the input image to modify the trimap, followed by alpha matte generation and propagation.

On another front, Lu et al. [34] introduced IndexNet, proposing indices-guided unpooling in the decoder as a better alternative to traditional upsampling. This new approach allowed for better retention of boundary details, outperforming DIM [6] using a lighter MobileNetV2 backbone [35].

Forte and Pitié [36] introduced a low-cost modification in F,B, α (FBA) Matting, allowing matting networks to predict foreground and background colors along with the alpha matte. This method not only achieved state-of-the-art results but was also efficient in terms of computational and memory costs.

Lastly, a notable study was done on trimap-based image matting techniques. In this study, the methodology encompassed a comprehensive three-part process involving segmentation, trimap generation, and matting. The segmentation phase utilized the strengths of UNet and UNet++ models, both built upon the EfficientNet-B4 backbone and pre-trained on the expansive ImageNet dataset. These models were later fine-tuned with data from the P3M-10K dataset, which provided high-quality inferred ground truth segmentation.

Network	P3M-500-P		P3M-500-NP	
	MSE	MAD	MSE	MAD
DIM(Their, UNet)	0.012	0.016	0.013	0.018
DIM(Their, UNet++)	0.010	0.013	0.015	0.011
DIM(Their, UNet++, Fine-Tuned)	0.010	0.013	0.015	0.011
IndexNet(Their, UNet)	0.012	0.015	0.014	0.018
IndexNet(Their, UNet++)	0.009	0.013	0.011	0.015
FBA Matting(Their, UNet)	0.012	0.015	0.014	0.017
FBA Matting(Their, UNet++)	0.009	0.012	0.011	0.014

Table 3.1: Comparative Analysis of various Network Combinations for Image Matting. This table is taken from [37].

In the trimap generation stage, an encoder-decoder network was trained on P3M-10K's precise ground truth trimaps. The matting phase was conducted using three advanced models: DIM, FBA Matting, and IndexNet, each initially pre-trained on the Adobe Deep Matting dataset. The FBA Matting model, in particular, received additional optimizations to improve data augmentation, enhancing its performance.

The standout performance of the FBA Matting model when integrated with the UNet++ segmentation network was evident, particularly in terms of the mean absolute difference (MAD) metric on both P3M-500-P and P3M-500-NP benchmarks. This system takes an image $\mathbf{I}_s$ as an input in $\mathbb{R}^{H\times W\times 3}$ and outputs the alpha matte α_s in $\mathbb{R}^{H\times W\times 1}$.

Trimap-Free methods

The progress of image matting entered a new phase with the development of trimap-free methods. Earlier techniques relied heavily on trimaps —- additional inputs that were made by users. To address this, researchers shifted their focus towards eliminating the need for trimaps, which included exploring automated trimap generation [23, 38, 39].

One of the leading developments in trimap-free matting was made by Chen et al. with their Semantic Human Matting (SHM) method [23]. They came up with a creative way of using a semantic segmentation network to create a trimap automatically. After

generating the trimap, they used another matting network to process it. The idea behind using two networks was to get a wide understanding of the scene as well as capture fine details. This approach helped in accurately figuring out the alpha matte, which is crucial for accurate matting results.

Following a similar path, Zhang et al. [40] introduced the Late Fusion CNN for Digital Matting (LF). They used a unique approach involving a single encoder that splits into two separate decoders for foreground and background separation. The key innovation of their method was how they combined the outputs from these decoders for aiming to refine the alpha matte predictions. Their method demonstrated promising results on a specialized benchmark for matting.

The evolution of trimap-free matting took a significant leap forward with the introduction of innovative methods like Glance and Focus Matting (GFM) and P3M-Net by Li et al. [41, 42] These methods stand out due to their multi-task frameworks and the incorporation of privacy considerations during the training phase. The development of new datasets, such as P3M-10k has been instrumental in enhancing the capability to thoroughly evaluate and compare various matting techniques. This progress has opened up more opportunities for detailed and reliable assessments in the field

This progression towards trimap-free image matting reflects the continuous effort in the field to make user interaction simpler while either maintaining or enhancing the quality of matting results. Each advancement, whether it be in automated trimap generation or fully trimap-free solutions, represents a significant step towards making image matting more accessible and efficient.

3.3 Video Matting

The shift from focusing on image matting to video matting was a major change in the field of computer vision, introducing new challenges and methodologies. This change

happened by the added complexity of the temporal dimension in videos. Unlike image matting, which deals with static scenes, video matting requires addressing the changing scenes over time. This necessity led to the adaptation and development of new methodologies in video matting. In response to this shift, several notable research works came out. These works concentrated on understanding and using the movement and continuity in videos to improve matting techniques.

In the following sections, the discussion will center around the research developments in the field of video matting and provide a complete overview of the progression from traditional approaches to deep learning-based methods.

3.3.1 Traditional Methods

The journey into using temporal propagation for video matting started with an important study by Apostoloff et al. [43] They brought a new perspective by using a Bayesian approach, which was based on learned image priors. Their innovative method used a Markov Random Field (MRF), a mathematical model, to understand the connections between matting elements over time and across different frames in a video. This approach was particularly effective in dealing with small movements in videos and achieved high-quality results in video matting.

Following the foundational work in temporal propagation for video matting, Choi et al. [44] made a significant contribution with their method that utilized multiple frames to enhance matting results. Their approach integrated both spatial and temporal information, aiming to achieve a more accurate and refined outcome in video matting. This method represented a considerable step forward in effectively utilizing the temporal continuity that is a natural aspect of videos.

The development of video matting techniques continued to evolve with the significant work of Li et al. [45] They introduced a motion-aware approach in their study, specifically designed to handle substantial motion in video sequences. Their methodology

cleverly integrated motion information into the matting Laplacian, leading to improved performance in videos characterized by noticeable motion.

In a similar vein, Lee et al. [46] made a noteworthy contribution with their research on Temporally Coherent Video Matting. They addressed the issue of flickering around the foreground boundaries, a common problem arising from determining alpha mattes frame by frame. Their work underscored the importance of a more robust method that ensures consistency over time in video matting.

3.3.2 Deep Learning-Based Methods

With the introduction of deep learning into video matting, the field has undergone a significant transformation. This shift to deep learning-driven methods has resulted in far more efficient and accurate techniques compared to traditional approaches. This advancement has been possible due to improved computing power and the availability of large datasets, which have enabled the training of more complex and refined models. This transition highlights a major development in the field, leveraging deep learning's powerful capabilities to effectively tackle the challenges of video matting.

Trimap-Based methods

Zhang et al. [47] made notable strides in deep learning-based video matting by focusing on achieving temporally coherent results . They developed an attention-based temporal aggregation module to improve video matting networks. This module works by calculating temporal correlations for pixels in feature space, effectively addressing motion noises, a major challenge in video matting. An innovative aspect of their work was introducing a new loss term to train the attention weights, greatly enhancing video matting performance. This loss term is crucial for guiding the learning of attention weights, making the method more robust against common video matting challenges like compression artifacts, changes in appearance, and motion. They also approached the trimap generation

issue by adjusting a top-notch video object segmentation network with a small number of user-marked key frames. This technique efficiently created necessary trimaps for video matting, reducing the manual effort typically needed for such annotations. Moreover, they compiled an extensive video matting dataset consisting of 80 training and 28 validation foreground video clips with accurate alpha mattes. This dataset fills a critical gap in resources available for training and evaluating video matting methods, fostering the development of new approaches. Their experimental results showed that their method can produce high-quality alpha mattes, even in challenging situations, marking an improvement in temporal coherence and overall video matting performance.

Lin et al. [48] then introduced a groundbreaking approach for real-time, high-resolution human video matting. Their method uses a recurrent architecture, which is a significant change from the typical frame-by-frame processing, effectively utilizing temporal information. This approach not only improves temporal coherence and matting quality but is also more efficient and faster. At the heart of this method is a feature-extraction encoder inspired by the best semantic segmentation networks, using MobileNetV3-Large as its main structure. The recurrent decoder, which includes ConvGRU at multiple scales, is carefully chosen for its ability to handle both long-term and short-term temporal information efficiently. The design of this decoder, with its bottleneck block, upsampling blocks, and output block, effectively translates the final features into outputs that include alpha, foreground, and segmentation predictions. Additionally, Johnson et al. proposed a dual training strategy that focuses on both matting and semantic segmentation, aiming to make the model more robust and reduce the overfitting problems that are common in models trained on synthetic data. They also implemented the Deep Guided Filter (DGF) [49], which is particularly useful for high-resolution videos like 4K and HD, ensuring top-quality matting results.

Trimap-Free methods

The field of real-time video matting underwent a significant change with the work of Johnson et al. [50] and their innovative MODNet (Matting Objective Decomposition Network). Breaking away from the traditional dependence on green screens for alpha matte extraction, MODNet showcased the possibility of real-time portrait matting using just a single input image. The brilliance of this system lies in its approach to breaking down the matting objective into smaller goals, which are then optimized together through clear constraints. This aspect is particularly crucial in video matting, where it's important to keep consistency and accuracy across different frames. The use of the Efficient Atrous Spatial Pyramid Pooling (e-ASPP) module in MODNet is a key factor for combining features from different scales, essential for accurate semantic estimation in video content. Furthermore, MODNet's Self-Supervised Sub-Objectives Consistency (SOC) strategy addresses the issue of domain shift, a common challenge in trimap-free video matting methods, thus ensuring adaptability and effectiveness with real-world data.

Similarly, Sun et al. [51] made a substantial contribution to video matting by focusing on deep learning techniques. Their framework introduced a Spatio-Temporal Feature Aggregation Module (ST-FAM), skillful at producing alpha mattes that are coherent both spatially and temporally. By extracting features at different levels and using information from multiple frames without depending on optical flow, this module tackled key challenges in video matting: ensuring spatial and temporal coherence and minimizing the need for detailed trimap annotations in each frame. Their innovative use of a correlation layer to propagate trimaps across frames is a notable step towards reducing the need for extensive user inputs, making the method more practical. Additionally, they provided a comprehensive video matting dataset with ground truth alpha mattes, including 10 high-resolution real-world videos with dense human-annotated trimaps frame by frame. This dataset is crucial for quantitative assessments. Their extensive evaluations showed that their method significantly outperforms traditional image-based and video matting

approaches, especially in complex situations with fast-moving objects or intricate backgrounds.

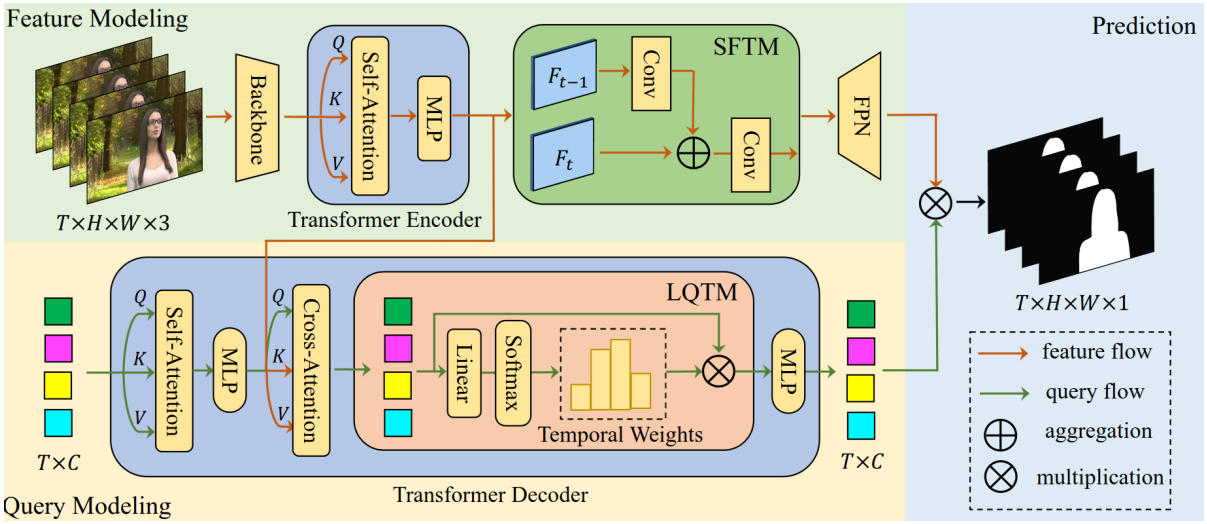


Figure 3.7: An overview of the VMFormer Architecture. This figure is taken from VMFormer[2].

Lastly, Li et al. [2] contributed VMFormer, an end to end video matting system. It takes a video sequence $\mathbf{I} = \{\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_T\}$ in $\mathbb{R}^{T\times H\times W\times 3}$ as input, where T is the number of frames. The output is the final prediction of alpha mattes $\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_T\}$ in $\mathbb{R}^{T\times H\times W\times 1}$. This model has two primary branches: the feature modeling branch and the query modeling branch. The feature modeling branch integrates a CNN-based backbone with a transformer encoder, adeptly extracting feature maps from the video sequences. This encoder is built upon multiple blocks, each containing a self-attention layer and a multi-layer perceptron. These components are crucial for achieving global modeling across each frame of the video sequence.

The query modeling branch of VMFormer employs a transformer decoder, featuring a cross-attention module. This module enables learnable queries to interact with the entire feature sequences. The significance of this interaction lies in its capacity to facilitate these queries in learning global representations, thereby enhancing the accuracy of alpha matte predictions.

For temporal aspects, VMFormer introduces two innovative modules: the Long-Range Query-based Temporal Modeling (LQTM) and the Short-Range Feature-based Temporal Modeling (SFTM). LQTM leverages self-attention mechanisms to apply temporal modeling based on queries, efficiently learning temporal weights. SFTM, conversely, focuses on short-range temporal modeling by recurrently aggregating successive feature maps.

In the final stage of the VMFormer model, alpha mattes are predicted by combining the largest feature map with queries from the transformer decoder, utilizing batch matrix multiplication. This final step is pivotal in generating the accurate and detailed alpha mattes essential for video matting applications.

In this chapter, we reviewed a range of significant works in video matting, highlighting the field's progress and the challenges distinct to it, especially when compared to image matting. One of the key challenges in video matting is dealing with temporal information between frames, making these algorithms more complex and demanding in terms of resources than those for static images.

In our research, we are exploring a novel approach that seeks to extend the principles of image matting to video matting. Rather than developing new video matting algorithms from scratch, our focus is on refining the alpha mattes generated by existing matting systems.

Chapter 4

A New Approach to Solve Video Matting

In this chapter will detail our methodology, concentrating on how we aim to improve the initial alpha mattes generated by matting systems. We will discuss the integration of advanced segmentation models and meta-learning techniques in our process.

4.1 Problem Statement

As we discussed earlier, video matting is a process where the goal is to separate an object from its background in video frames. It involves calculating the transparency level of each pixel in each frame of a video, known as an alpha matte, which is challenging but crucial for creating realistic scenes.

Considering a video sequence denoted as $\mathbf{I} = \{\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_T\}$, each frame I_i in this sequence can be modeled as a composite of a foreground image F_i and a background image B_i, merged using an alpha matte α_i within the range [0, 1]:

$$I_i = \alpha_i F_i + (1 - \alpha_i) B_i, \tag{4.1}$$

Here, α_i is the alpha matte for each frame. The objective of video matting is to accurately predict the sequence of alpha mattes $\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_T\}$ for all the frames. This task is inherently challenging due to the under-constrained nature of the problem: for each pixel, there are seven unknown variables - three from F_i, three from B_i, and one from α_i - compared to only three known values from the RGB color channels in I_i.

In our research, we aim to address the unique challenges of video matting by adopting an approach that extends the principles of image matting to the dynamic realm of video sequences. This method is predicated on the concept of transfer learning, where techniques successful in one domain are adapted and applied to another. Our primary objective is to enhance the accuracy and quality of alpha mattes $\{\alpha_1, \alpha_2, \ldots, \alpha_T\}$ in video matting, which are crucial in determining the realism and effectiveness of the final video content. To achieve this, we introduce two concepts in our methodology: Adapters and Backbones.

- **Backbones:** In this research, a backbone refers to the core algorithm or system that initially processes the video frames to generate initial alpha mattes. In our architecture design, the notation $\mathcal{B}$ is adopted to represent the backbone system.

- **Adapters:** To refine the output of the mentioned backbone system, we incorporate adapters. An adapter in our research is a segmentation model. The role of an adapter is to boost the quality of initial alpha mattes. In our architecture design, the adapter is denoted as $\mathcal{A}$.

4.2 Architecture Design

In this section, we will explore the architecture design that we chose for implementing our approach.

4.2.1 Boosting with Adapters (BwA)

In the **Boosting with Adapters** method, our video matting framework employs a two-part system to refine alpha mattes and enhance the accuracy of video matting results. The process begins with an input sequence that passes through a frozen backbone, denoted as B. This backbone is tasked with generating an initial alpha matte, α_t, from the input frames $\mathbf{I_t}, \mathbf{I_{t-1}}, \mathbf{I_{t-2}}, \ldots, \mathbf{I_{t-l}}$. The initial alpha matte represents the system's preliminary effort at distinguishing the foreground from the background in video frames.

Following the initial generation, both the alpha matte α_t and the original input undergo processing by an adapter A, which possesses trainable weights. The primary role of this adapter is to learn from the deficiencies of the frozen backbone B in producing the alpha matte. It fine-tunes its parameters specifically to address and correct these errors. By adapting to the mistakes made by the backbone, the adapter refines the initial alpha matte, outputting a more accurate and polished version.

In our approach, we intentionally avoid using a residual layer between the backbone B and the adapter A. This configuration is critical as it prevents the backbone's output from potentially overpowering the adapter's corrective actions, ensuring the adapter focuses solely on addressing specific errors in the alpha matte independently. This methodological choice significantly enhances the precision of the error correction process.

Managing the amount of temporal information processed is crucial to maintaining optimal model performance. Our system carefully balances this by incorporating an optimal amount of temporal data to effectively capture motion dynamics without unnecessarily complicating the matting process.

The rationale for freezing the backbone during the fine-tuning phase of the adapters is to reduce computational demands. By keeping the backbone's output consistent, we prevent it from affecting the learning trajectory of the adapters, allowing them to refine their performance based on a stable input consistently. This approach enhances system efficiency, particularly in practical applications where computational resources are

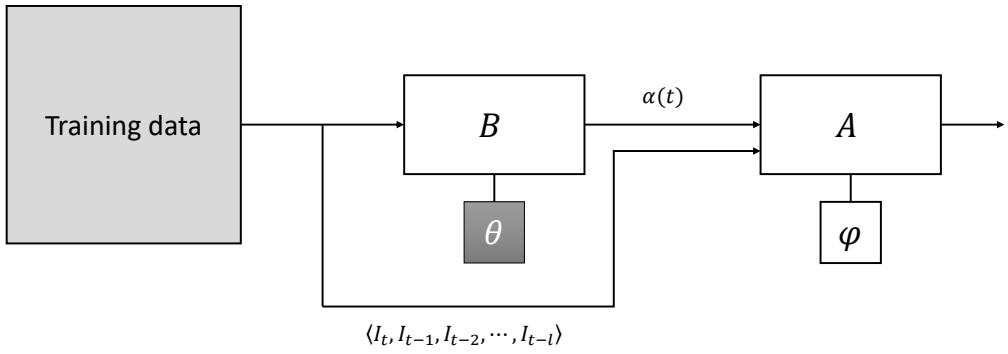


Figure 4.1: Workflow diagram highlighting the "Boosting with Adapters" approach, where the backbone system remains frozen and the adapters are fine-tuned with the purpose of alpha matte refinement. This figure demonstrates the process for both single and multi-frame scenarios, showing how the initial alpha matte generated by the frozen backbone, is subsequently refined by the fine-tuned adapters to produce the improved alpha matte.

limited.

These strategic design choices ensure our system not only achieves high accuracy in video matting but also remains efficient and adaptable, capable of handling the dynamic and complex nature of video content with minimal computational resources.

4.2.2 Expert Selection

Following the initial refinement phase in the **Boosting with Adapters** approach, we implement an **Expert Selection** phase. Given that adapters possess varying architectures, they exhibit different performances during the error correction process. To effectively manage this variability and optimize the system performance, we employ a subset of our training data, referred to as the selection data, for this phase.

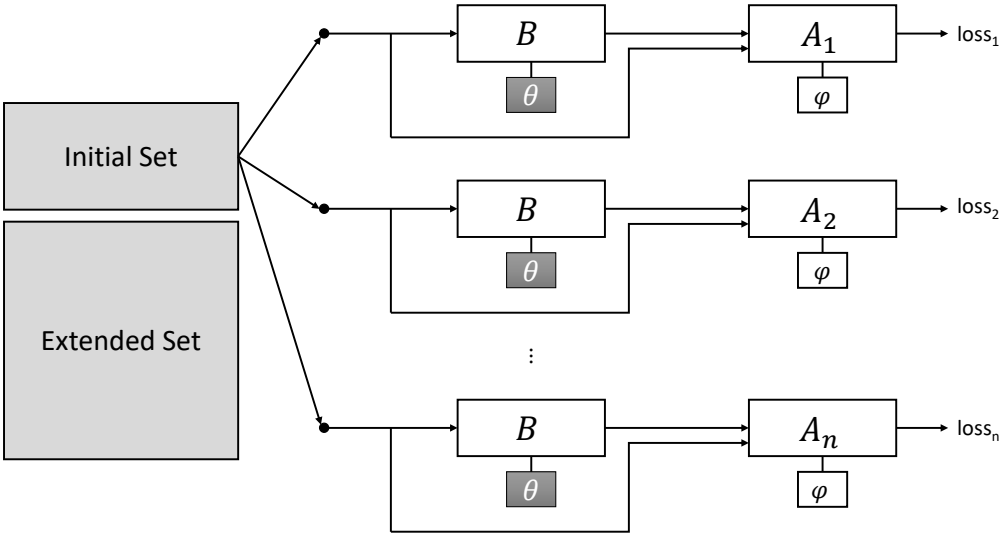


Figure 4.2: Diagram illustrating the Expert Selection phase. This figure shows the process of fine-tuning multiple adapters on a subset of training data, evaluating their performance based on specific metrics, and selecting the most effective adapters based on their performance on unseen test data. The selected adapters are then designated as expert adapters for subsequent phases, optimizing both computational resources and matting accuracy.

Each adapter $a_1, a_2, \ldots, a_n$ is fine-tuned within the framework established by the boosting process. after fine-tuning, these adapters are evaluated based on specific metrics, which will be detailed later in this chapter. The evaluation is conducted on unseen test data to ensure unbiased assessment of each adapter's performance.

The purpose of this **Expert Selection** phase is to identify 'expert adapters' that offer the most effective error correction for the given video matting task. This selection process is crucial as it allows us to harness the strengths of specific adapters without the need to deploy all within the operational environment, thus saving on computational costs. By fine-tuning and selecting experts on just a portion of the dataset, we can achieve significant efficiencies while maintaining high performance levels.

4.2.3 Boosting using Ensemble (BuE)

With the expert adapters identified in the previous phase, we proceed to implement the **Boosting Using Ensemble** approach. In this phase, the expert adapters $A_1^*, A_2^*, \ldots, A_n^*$ are now frozen with their weights fine-tuned from the Expert Selection phase. This setup ensures that the initial refinements each adapter learned are preserved and utilized in the subsequent processing steps.

The process begins with the input video frames passing through the frozen backbone B, which outputs an initial alpha matte. This alpha matte, along with the original input, is then processed by each of the frozen, fine-tuned expert adapters. Each adapter outputs its version of the refined alpha matte, reflecting its specialized adjustments.

Subsequently, these refined alpha mattes, together with the initial input, are fed into an ensemble layer h with trainable weights. The role of the ensemble layer is critical: it intelligently combines the outputs from each expert adapter. By learning the optimal way to integrate these multiple refined alpha mattes, layer h effectively enhances the overall matting result, pushing the accuracy and quality of the video matting even further.

The ensemble layer h can be designed in several ways, depending on the specific requirements of the application and the characteristics of the data. Common choices for the ensemble head include:

- **Weighted Average:** Where each expert's output is assigned a weight that reflects its relative importance, determined during the training process.

- **Concatenate and Convolving:** Where outputs from all experts are concatenated and passed through convolutional layers to integrate information.

- **Attention Mechanisms:** Which dynamically adjust the focus on different expert outputs depending on the input, allowing the model to adaptively prioritize more relevant features.

These choices each offer unique advantages in terms of flexibility, adaptability, and efficiency in integrating diverse expert opinions into a coherent output. Among the

options for designing the ensemble layer h, we opted for the Concatenate and Convolving method. This approach involves concatenating the outputs from all expert adapters, which are then processed through convolutional layers.

The concatenated output provides a comprehensive feature set that represents multiple perspectives on the matting task, which the convolutional layers can effectively synthesize. This method leverages the spatial processing capabilities of convolutional networks, making it possible to extract and refine features from combined inputs, thereby enhancing the overall matting accuracy. The convolutional layers act to distill and enhance the relevant features from the concatenated outputs, ensuring that the final matting result is both precise and robust to various video content challenges.

By using this design, our system gains the ability to effectively combine the strengths of each adapter, leading to improved performance and greater adaptability in handling complex video matting scenarios.

This ensemble approach leverages the strengths of multiple specialized adapters, blending their individual corrections into a cohesive and superior alpha matte. The decision to freeze the expert adapters' parameters during the fine-tuning of the ensemble layer is pivotal. It stabilizes the input to the ensemble layer, allowing it to focus solely on learning how to effectively combine the expert outputs without the complexity of changing individual adapter behaviors.

In addition, the number of expert adapters plays a crucial role in the efficiency and effectiveness of the ensemble. Utilizing too few experts can limit the diversity of error correction, potentially missing nuances in different video scenarios. Conversely, too many experts can lead to excessive gradient computations during training, increasing the complexity and computational cost of determining the relevance of each expert's contribution. This balance ensures that the ensemble method remains computationally efficient while maximizing the quality of the video matting output.

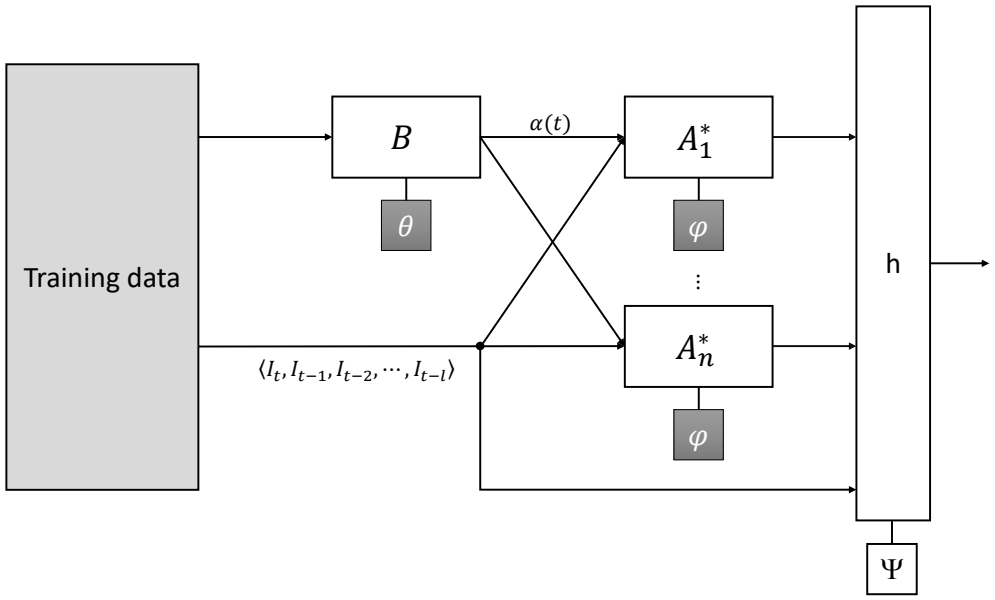


Figure 4.3: Diagram illustrating the "Boosting using Ensemble" method, where the backbone system and the fine-tuned adapters are frozen. This figure shows the integration of a convolutional neural network within our ensemble. The convolution head is crucial for analyzing and combining the outputs from the expert models and the backbone, leading to a sophisticated and refined final alpha matte for video frames.

4.3 Choices of Backbones

In our research, the selection of an appropriate matting system forms a crucial part of the architecture. This system, known as the backbone, is tasked with generating the initial alpha mattes, which are pivotal for our subsequent refinement process. The backbone we choose needs to possess the capability to process RGB images or a sequence of video frames, producing a grayscale alpha matte. A significant requirement for our chosen backbone is its ability to function autonomously, without necessitating additional inputs like a trimap, which is a common prerequisite in many matting systems. For the purpose of our study, we have selected two specific backbones, each distinguished by its specialized functionality:

- **FBAUNET++ [37]**: As we discussed in the related work, this system is primarily an image matting system and its efficacy lies in handling static images, making it an

excellent choice for scenarios where the primary focus is on single-frame processing. Since our main focus is on applying image matting systems to video data and expand from there, FBAUNET++ is a great base for our approaches.

- **VMFormer** [2]: While the main focus of our research is centered on adapting image matting techniques to video data, we have strategically incorporated VMFormer, a state-of-the-art video matting system, as another backbone for further exploration. This choice is detailed in the related work section. Using this system as our backbone allows us to investigate whether our methods can enhance even the most advanced video matting systems. This aspect of the research is crucial in assessing the potential of our methodologies to contribute improvements to existing cutting-edge video matting technologies.

4.4 Choices of Adapters

In our video matting process, the adapters, chosen from a range of segmentation models, play a crucial role beyond their traditional segmentation tasks. These adapters, integral to our methodology, are specifically employed to refine alpha masks. The refinement of alpha masks is a key step in enhancing the quality and accuracy of the alpha mattes initially produced by our selected backbones. For our research, we have carefully selected the following six segmentation models to serve as adapters, each offering unique capabilities: FPN [27], PAN [25], DeepLabV3 [52], UNet [28], LinkNet [26], MANet [24]. It's important to note that all these adapters are built upon the ResNet50 [53] architecture, utilizing its backbone structure with pre-trained ImageNet [54] weights.

4.5 Evaluation Metrics

We measure our performance with four metrics: Mean Absolute Differences (MAD), Mean Squared Errors (MSE), Gradient (GRAD), and Connectivity (CONN). MAD and MSE are universally recognized metrics in both image and video matting evaluations, providing a solid foundation for assessing the basic accuracy of the matting results.

However, for tasks involving motion, such as video matting, GRAD and CONN metrics are particularly crucial. These metrics are computed across multiple frames, rather than on a single frame basis, which makes them indispensable for evaluating the overall quality of the output video.

4.5.1 Mean Absolute Difference

The Mean Absolute Difference (MAD) is a statistical measure often used to quantify the error between the computed alpha matte and its ground truth. MAD is particularly useful for assessing the performance of matting algorithms as it provides a straightforward and intuitive measure of the average magnitude of errors. The formula for MAD is:

$$\text{MAD} = \frac{1}{N}\sum_{i=1}^{N}|\alpha_i - \alpha_i^*|. \tag{4.2}$$

Here, α_i represents the alpha value of the i^{th} pixel in the computed alpha matte, α_i^* is the alpha value of the i^{th} pixel in the ground truth, and N is the total number of pixels. The absolute difference $|\alpha_i - \alpha_i^*|$ is summed over all pixels and then averaged.

A lower MAD indicates that on average, the computed alpha values closely match the ground truth. This suggests that the matting algorithm is both accurate (close to the true value) and precise (having less variability). Also , In the context of image composition, a lower MAD means that the transitions between foreground and background are more seamlessly reproduced, leading to a more visually consistent and realistic composite image.

4.5.2 Mean Squared Error

Mean Squared Error (MSE) is another critical metric for evaluating the performance of alpha matting algorithms. It measures the average of the squares of the errors between the computed and ground truth alpha values, providing insight into the variance of the errors. The MSE is defined as:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\alpha_i - \alpha_i^*)^2. \tag{4.3}$$

In this formula, like what we had in MAD, α_i and α_i^* are the alpha values for the computed and ground truth mattes at pixel i, respectively. The squared difference $(\alpha_i - \alpha_i^*)^2$ is then averaged over all N pixels.

MSE gives more weight to larger errors due to the squaring of the differences. This means a lower MSE indicates not only general accuracy but also that there are no significant errors in any part of the matte. In addition to this, in alpha matting, edges and fine details are critical. A lower MSE value suggests that these areas are handled well, with fewer large discrepancies in transparency levels.

4.5.3 Gradient

Next error metric is the gradient error [55] which quantifies the difference in the rate of change in pixel intensities between the computed and ground truth alpha mattes. The gradient error is given by the equation:

$$\text{Gradient Error} = \sum_{i}(\nabla\alpha_i - \nabla\alpha_i^*)^q, \tag{4.4}$$

where i is the pixel index, $\nabla\alpha_i$ and $\nabla\alpha_i^*$ are the normalized gradients of the computed and ground truth alpha mattes at pixel i, respectively. The exponent q emphasizes larger discrepancies. Lower gradient error values indicate a close match between the computed

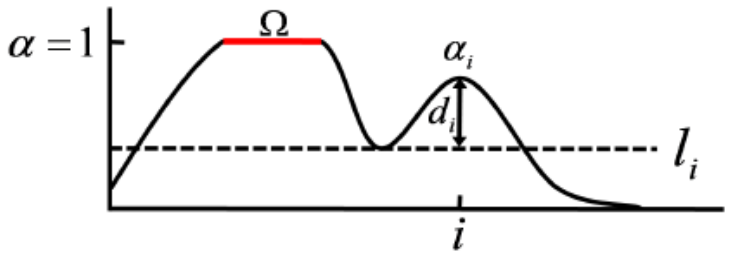


Figure 4.4: Connectivity error. This figure is taken from [55].

alpha matte and the ground truth. This close match ensures the accuracy of transitions and edges, important for the visual integrity of composite images. A lower error reduces visual mistakes, leading to a more realistic representation.

4.5.4 Connectivity

Lastly the concept of connectivity directly influences the quality of the output. Connectivity refers to the way pixels with similar transparency levels are grouped or connected in an image. The primary goal in evaluating connectivity is to ensure that the transition between the foreground and background in an image is seamless and visually coherent. To simplify the connectivity analysis, alpha mattes are often converted into binary images. This process involves classifying pixels as either part of the foreground or the background based on a predefined threshold. This binary representation makes it easier to evaluate the connectivity of pixels. The connectivity error can be defined as:

$$\text{Connectivity Error} = \sum_i (\phi(\alpha_i, \Omega) - \phi(\alpha_i^*, \Omega))^p. \tag{4.5}$$

Here, α_i is the transparency level of a pixel in the computed alpha matte, α_i^* is the transparency level in the ground truth matte, and Ω represents a source region, typically the largest contiguous opaque area. The function ϕ measures the degree of connectivity for a pixel, and the exponent p is usually set to 2 to emphasize larger differences.

The connectivity measure ϕ for a pixel is determined based on the distance $d_i = \alpha_i - l_i$,

where l_i is a threshold level that defines whether a pixel is connected to the source region Ω. The formula for ϕ is:

$$\phi(\alpha_i, \Omega) = 1 - (\lambda_i \cdot \delta(d_i \geq \theta) \cdot d_i). \tag{4.6}$$

In this expression, $\delta(d_i \geq \theta)$ is a function that ignores small variations in d_i below a threshold θ, focusing on significant connectivity differences. λ_i is a weighting factor that accounts for the average distance of disconnected pixels from Ω.

A lower connectivity error indicates that the alpha matte accurately represents the connectedness of pixels, closely resembling the ground truth. This is crucial for ensuring that the foreground and background are correctly segmented, leading to a more natural-looking image. An image with a lower connectivity error has fewer disconnected regions. This is especially important in maintaining the integrity of edges and transitions between different regions in an image. In image compositing, where alpha mattes play a key role in blending images, excellent connectivity ensures smooth and natural transitions.

By incorporating all four metrics, with a particular emphasis on GRAD and CONN, our evaluation strategy addresses both the static accuracy of individual frames and the dynamic integrity of the entire video sequence. This approach allows us to thoroughly assess the effectiveness of our video matting techniques, ensuring high-quality results across diverse video content.

Chapter 5

Experiments and Result

5.1 Datasets

In this section, we introduce the datasets that form the basis of our video matting research. The selection of appropriate datasets is crucial for our study, and for this purpose, we have chosen two datasets: VideoMatte240K[56] and BG-20k[57]. These datasets provide a wide range of data necessary for the development and evaluation of our video matting methods.

5.1.1 Datasets Overview

VideoMatte240K: The primary dataset for our video content is VideoMatte240K, developed by Lin et al. This dataset comprises 484 green screen videos which have been converted into 240,709 frames. Each frame includes a corresponding alpha matte and foreground element. The majority of these frames are in 4K resolution, with a portion in HD. The dataset encompasses a diverse array of human subjects, various clothing styles, and a multitude of poses. This diversity is essential for testing the effectiveness and adaptability of our video matting models across different scenarios.

BG-20k: For background images, we utilized the BG-20k dataset provided by Li et al. This dataset consists of 20,000 high-resolution images, carefully selected to ensure the absence of dominant foreground objects. The diversity of the backgrounds, ranging from natural settings to urban environments, is advantageous for creating synthetic test scenarios. This allows for a comprehensive assessment of our matting models against a variety of background conditions.

5.1.2 Background Shuffling

For each frame extracted from the VideoMatte240K dataset, we randomly assigned a background from the BG-20k dataset. This meant that no two frames had the same background, even if they originated from the same video clip. By doing this, we created a training dataset where each frame presented a unique combination of foreground subject and background scene.

The reason behind this approach was to encourage the model to focus more on the foreground elements, irrespective of the background. In real-world scenarios, the background can vary significantly, and our model needs to reliably separate the foreground from a different potential backgrounds. By training the model on frames with constantly changing backgrounds, we aimed to reduce the model's dependency on background features and enhance its ability to isolate the foreground.

We anticipated that this method would lead to the development of a more robust matting model. A model trained in such a diverse and unpredictable environment is expected to perform more consistently and accurately, irrespective of the background conditions in the input frames.

5.1.3 Dataset Categorization

For our study, we organized the datasets into two main categories, each serving a specific purpose in our experiments. These are the 'Initial Set' and the 'Extended Set'.

Figure 5.1: Showcasing Background Shuffling in Dataset: This figure consists of four frames from our dataset, illustrating the technique of background shuffling. Each frame features the same foreground subject from one video, set against a different background.

Initial Set: For the initial stage of our experimental work, we utilized a dataset named the 'Initial Set', comprising 1,000 frames. This set includes 200 frames from each of five different videos. These videos were selected to encompass a wide variety of scenarios, ensuring a diverse range of data for our analysis. The main purpose of using this set is to conduct fine-tuning of our adapters. This phase was crucial in identifying the most effective adapters, termed as 'experts', for use in our ensemble method, thereby laying the groundwork for the subsequent stages of our research.

Extended Set: Our research also incorporates a more extensive dataset, known as the 'Extended Set', which includes 10,000 frames. This dataset is composed of 100 frames from each of 100 diverse videos from our dataset. The Extended Set, encompassing a wide variety of human activities and scenarios, served as a robust platform for evaluating the performance of our models on a more extensive scale.

5.2 Experimental Setup

The training process was conducted using images with a resolution of 480 x 640 pixels, and the computations were performed on an NVIDIA V100 GPU. We utilized momentum Stochastic Gradient Descent for updating the model parameters, with a learning rate set at 0.0001. The training was done over 50 epochs, and to prevent overfitting, an early stopping criterion was implemented, terminating the training if no improvement was observed for 3 consecutive epochs. The batch size during training was maintained at 16,

a size chosen to optimize the learning process within the computational constraints of our setup. Additionally, we employed Dice Loss [58] as the loss function for the training, aligning it with the specific requirements and objectives of our matting tasks.

5.3 Experiments and Result

In this section, we will explore our experiments with FBAUNet++ and VMFormer as our backbone, focusing on refining video data matting. The experiments are structured in stages, utilizing our approaches, Boosting with Adapters and Boosting using Ensemble. We evaluated these methods using metrics such as MAD, MSE, GRAD, and CONN, to assess the effectiveness of various adapters and training approaches.

5.3.1 FBAUNet++ as the Backbone

Following our motivation to apply image matting systems to video data, we selected FBAUNet++ as our backbone for the first stage of our experiments. The first stage of our experiments was centered around refining alpha mattes using our BwA approach. We fine-tuned six different adapters with our initial training set. The primary objective in this phase was to improve the quality of the alpha mattes and assess the capability of each adapter in enhancing the matting output using this approach. After completing the fine-tuning process, we evaluated the performance of all six adapters. This evaluation was based on their effectiveness in refining alpha mattes and was conducted using an unseen test set to ensure unbiased results.

Based on the results presented in Table 5.1, the integration of adapters with the FBAUNet++ backbone generally enhances matting performance across several metrics—MAD, MSE, GRAD, and CONN. Notably, when utilizing two input frames, the performance across all adapters improves significantly compared to configurations with either one or four frames. This improvement indicates an optimal balance of temporal

Method		#Input Frames	MAD↓	MSE↓	GRAD↓	CONN↓
Backbone	Adapter					
FBAUNet++	None	1	33.95	33.76	11.83	11.17
FBAUNet++	FPN	1	35.45	31.04	20.91	9.88
		2	**13.18**	**8.77**	13.15	**3.04**
		4	33.31	28.89	25.75	9.26
FBAUNet++	Unet	1	22.45	18.04	13.51	5.89
		2	15.45	10.04	**11.15**	3.34
		4	22.08	17.67	14.92	5.78
FBAUNet++	PAN	1	51.12	46.71	23.87	14.67
		2	17.27	12.86	14.47	4.28
		4	31.66	27.25	30.18	8.75
FBAUNet++	MANet	1	33.15	28.74	17.65	9.15
		2	23.87	19.46	13.53	6.33
		4	39.85	33.64	31.49	12.74
FBAUNet++	LinkNet	1	30.48	26.07	15.49	8.35
		2	17.92	13.51	13.87	4.53
		4	65.49	61.08	22.91	19.25
FBAUNet++	DeepLabV3	1	19.77	15.36	15.98	5.05
		2	16.08	11.67	14.43	3.93
		4	31.48	27.07	27.34	8.71

Table 5.1: Effectiveness of BwA Approach in Matting Performance with FBAUNet++ Backbone and Various Adapters. This table compares the matting performance metrics – MAD, MSE, GRAD, and CONN – for various adapters incorporated within the FBAUNet++ backbone using our BwA method. Performance is evaluated over 1, 2, and 4 input frames, with lower metric values indicating improved matting quality.

information, which is sufficient to capture relevant motion dynamics without introducing excessive noise or complexity that could degrade the adapter's performance.

The optimal performance with two frames suggests that while the base models are designed to process single images, their capability in video matting can be substantially improved by incorporating a moderate amount of temporal data. This supports the decision to employ temporal information carefully in the video matting process, ensuring enough context for effective matting without overwhelming the computational framework.

Moreover, the selection of FPN and UNet as expert adapters was based on their

Method		Training Set	MAD↓	MSE↓	GRAD↓	CONN↓
Backbone	Adapter					
FBAUNet++	None	—	33.95	33.76	11.83	11.17
FBAUNet++	FPN	Initial	13.18	8.77	13.15	3.04
		Extended	**9.68**	**5.27**	3.52	**1.96**
FBAUNet++	Unet	Initial	15.45	10.04	11.15	3.34
		Extended	10.29	5.88	**2.78**	2.14

Table 5.2: Impact of Training Data Volume on Adapter Performance in the BwA approach. This table presents a comparison of FPN and UNet adapters' performance on the FBAUNet++ backbone, contrasting results with initial and extended training data sets.

consistent superior performance in comparison to other adapters. These models not only demonstrated the best matting quality but also showed robustness across different metrics, validating their selection for further experiments. Setting the number of input frames to two for subsequent tests was a strategic choice, derived from empirical evidence of enhanced performance, ensuring that further evaluations and optimizations are grounded in the most effective configuration observed.

Having selected our expert adapters, we moved on to the next phase of our experiment with the aim of determining whether a larger volume of training data could further enhance the performance of these experts. The results of this experiment, as detailed in Table 5.2, clearly indicate a positive impact from the increased data volume on the performance of our expert adapters. This improvement underscores the fact that a more extensive dataset contributes significantly to the refinement process, ultimately leading to enhanced matting accuracy.

In this phase of our research, we demonstrated the effectiveness of meta-learning techniques in enhancing video matting performance through the implementation of an ensemble method, specifically the BuE approach. By freezing the weights of our expert adapters, which had been fine-tuned on an extended training set, we ensured that their refined capabilities were fully leveraged. The crucial role was then assigned to the con-

Method			MAD↓	MSE↓	GRAD↓	CONN↓
Type	Backbone	Adapter				
Base	FBAUNet++	None	33.95	33.76	11.83	11.17
BwA	FBAUNet++	FPN	9.68	5.27	3.52	1.96
	FBAUNet++	UNet	10.29	5.88	2.78	2.14
BuE	FBAUNet++	FPN + UNet	**8.37**	**5.16**	**1.33**	**0.98**

Table 5.3: Enhanced Matting Performance through BuE approach. The table presents a comparison of matting performance metrics – MAD, MSE, GRAD, and CONN – highlighting the effectiveness of the ensemble technique.

volution head within our ensemble architecture. This component was responsible for synthesizing the inputs from the frozen expert adapters and effectively determining the final output of the ensemble system. This setup showcases the potential of meta-learning to optimize video matting processes, highlighting the convolution head's pivotal role in refining and integrating the contributions of each adapter in the ensemble.

The results from the BuE approach, as detailed in Table 5.3, substantiate a significant finding: the combined performance of the two expert adapters, FPN and UNet, within the ensemble surpasses their individual performances. This outcome is particularly notable in the metrics for Mean Absolute Difference (MAD), Mean Squared Error (MSE), Gradient Magnitude Error (GRAD), and Connectivity Error (CONN), where the ensemble method achieves the lowest values across all metrics.

5.3.2 VMFormer as the Backbone

For the final stage of our study, we aimed to test the universality and effectiveness of our meta-learning approaches, namely Boosting with Adapters (BwA) and Boosting using Ensemble (BuE), by implementing them with a different backbone, VMFormer. This stage was crucial for demonstrating the flexibility of our methods and their capability to enhance even advanced video matting systems like VMFormer.

The results from Table 5.4 demonstrate that while the Boosting using Ensemble

Method			MAD↓	MSE↓	GRAD↓	CONN↓
Type	Backbone	Adapter				
Base	VMFormer	None	**6.39**	**1.51**	1.05	0.41
BwA	VMFormer	FPN	7.21	2.46	0.68	0.58
	VMFormer	UNet	7.14	2.53	0.59	0.57
BuE	VMFormer	FPN + UNet	6.51	2.18	0.57	0.39

Table 5.4: Performance Comparison of Base, BwA, and BuE Methods when VMFormer is the Backbone. This table delineates the matting performance metrics – MAD, MSE, GRAD, and CONN – for the VMFormer backbone using various approaches. The results provide insights into the effectiveness of each method in refining the quality of video matting.

method improved metrics such as GRAD and CONN, the gains in Mean Absolute Difference (MAD) and Mean Squared Error (MSE) were marginal. This is particularly significant given that VMFormer is already optimized to minimize errors on these global metrics, suggesting that our approaches encounter limitations in enhancing performance where the baseline system exhibits minimal errors, particularly in very detailed aspects of the matting process.

These findings highlight a critical limitation: our system struggles to further refine performance in areas where errors are already reduced to very fine details, which are less detectable and harder to correct. VMFormer's state-of-the-art capabilities mean that most significant errors have already been addressed, and the remaining inaccuracies often occur in intricate details that are challenging to capture and improve upon further with the current meta-learning strategies.

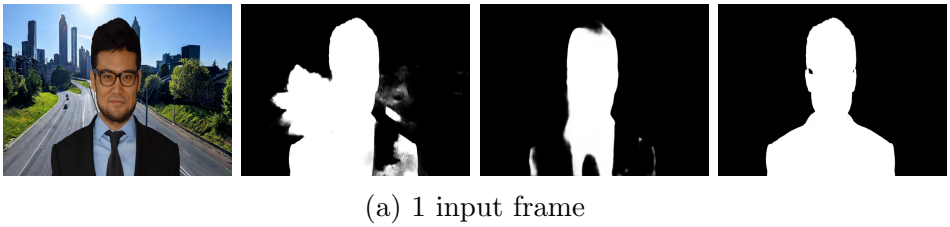
This limitation underscores the importance of developing new techniques or refining existing ones that can address these minute details more effectively. It also highlights the need for ongoing research into enhancing the adaptability and resolution of meta-learning models, particularly in high-performance contexts where improvements require handling subtleties that standard approaches may not adequately resolve.

5.4 Qualitative Analysis

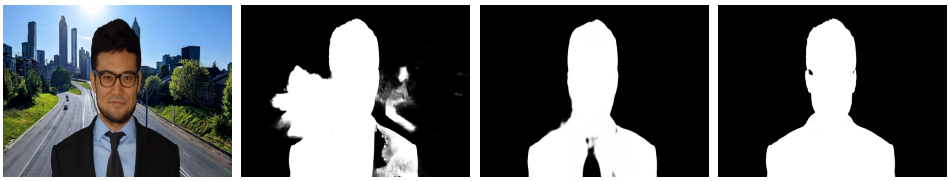
Transitioning from the quantitative results of our experiments, we now shift our focus to a qualitative analysis, providing a more visual and intuitive understanding of our methodologies' effectiveness. This section presents a series of figures that visually demonstrate the alpha matte refinement process achieved through our different approaches.

Figure 5.2 shows the qualitative results of different adapters in the Boosting with Adapters Approach using the FBAUNet++ backbone. This visualization helps in understanding the enhancements provided by various adapter configurations. Figure 5.3 illustrates the qualitative impact of the number of input frames on the Boosting with Adapters Approach. The figure highlights how the input frame count affects the refinement outcomes. Figure 5.4 demonstrates the impact of training data size on the Boosting with Adapters Approach, specifically using the UNet with FBAUNet++ backbone. This figure provides insights into the scalability and effectiveness of the approach with different data volumes. Lastly, Figure 5.5 compares the Boosting with Adapters approach to the Boosting using Ensemble Approaches. This comparison is crucial for understanding the relative advantages and limitations of each method.

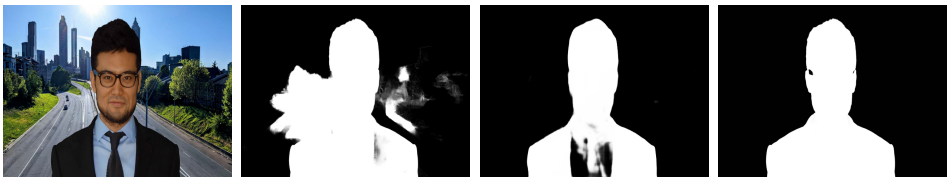
Figure 5.2: Qualitative Results of Different Adapters in Boosting with Adapters Approach using FBAUNet++ Backbone: This figure provides a qualitative visualization of the results achieved with various adapters as part of the 'Boosting with Adapters' approach, complementing the FBAUNet++ backbone. Displayed from left to right in each sub-figure are the previous frame, the current frame, the initial alpha matte produced by FBAUNet++, the refined alpha matte by each adapter, and the ground truth alpha matte. Arranged from top to bottom, the sub-figures offer a comparative qualitative analysis of the adapters - FPN, UNet, PAN, MANet, LinkNet, and DeepLabV3.



(a) 1 input frame



(b) 2 input frame



(c) 4 input frame

Figure 5.3: Qualitative Impact of Number of Input Frames on Boosting with Adapters Approach. This figure presents the performance of BwA approach where the backbone is FBAUNet++ and the adapter is FPN model for 1, 2, and 4 input frames. From left to right, each sub-figure shows the last frame of the input sequence, the initial alpha matte from FBAUNet++, the refined alpha by FPN, and the ground truth.

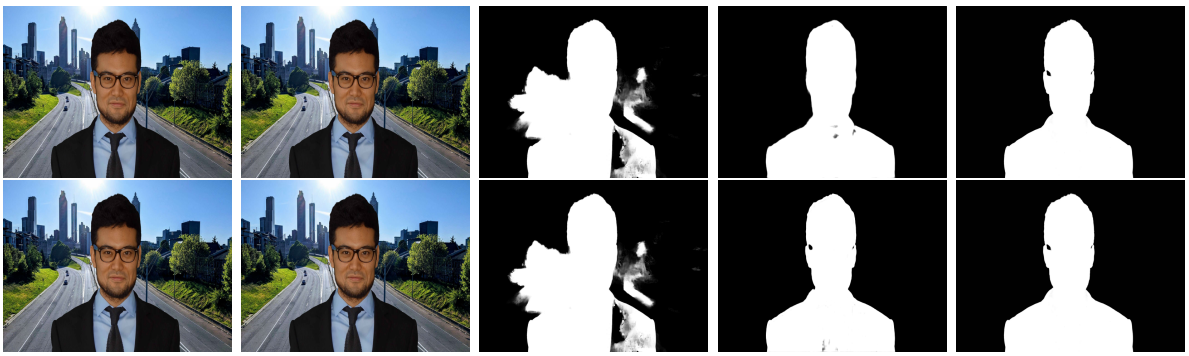


Figure 5.4: Impact of Training Data Size on Boosting with Adapters Approach: UNet with FBAUNet++ Backbone. This figure provides a qualitative comparison of the UNet adapter's performance in the BwA approach, emphasizing how varying sizes of training data influence matting quality. The top row illustrates the results of UNet fine-tuned with the initial, smaller training set, while the bottom row shows its performance when trained with the more extensive, extended set. In each sub-figure, displayed from left to right, are the previous frame, the current frame, the initial alpha matte generated by FBAUNet++, the refined alpha matte by UNet, and the ground truth alpha matte.

Figure 5.5: Comparison of Boosting with Adapters and Boosting using Ensemble Approaches. This figure offers a side-by-side comparison showcasing the outcomes of the BwA and BuE methods. The first row shows the results of the FPN adapter employed in the BwA approach, while the second row shows the UNet adapter's performance within the same approach. The third row presents the results of the BuE approach, where both FPN and UNet adapters are used in a frozen state as part of the ensemble. In each row, from left to right, the sequence includes the previous frame, the current frame, the initial alpha matte produced by the FBAUNet++ backbone, the refined alpha matte from the respective approach, and the ground truth.

Chapter 6

Conclusion

6.1 Contributions

This thesis has made several contributions to the field of video matting, which are outlined as follows:

- **Adaptation of Image Matting to Video Contexts:** We have successfully extended image matting techniques to video matting by applying principles of transfer learning. This approach specifically addresses the dynamic complexities inherent in video, such as motion and temporal continuity, enhancing the adaptability of matting techniques to the unique challenges of video sequences.

- **Development of Boosting with Adapters (BwA):** Our research introduces a novel boosting method that utilizes advanced segmentation models to refine alpha mattes. This method significantly improves the accuracy and realism of video matting results by correcting the mistakes of the matting process, thereby producing higher quality mattes.

- **Development of Boosting using Ensemble (BuE):** We developed an ensemble method that combines the strengths of multiple fine-tuned adapters. This strat-

egy not only enhances the overall quality of video matting but also increases the robustness and effectiveness of the matting systems across varied video content.

- **Multi-stage training** In our approach, we designed a two-stage process to enhance video matting. Initially, we fine-tuned adapters and selected the most effective ones based on their performance. Subsequently, we froze these adapters' weights and trained an ensemble layer to intelligently combine their outputs, significantly improving the overall quality of the matting results.

6.2 Limitations

While our research has advanced the field of video matting, it is not without limitations:

- The computational resources required for fine-tuning and training the models are substantial, which may limit their applicability in resource-constrained environments.

- Our models have shown promising results on the datasets used; however, their generalization across a broader range of video scenarios remains a challenge. This includes varying conditions such as lighting, object complexity, and motion dynamics.

- The complexity introduced by the ensemble method, although effective in enhancing performance, requires careful management to maintain a balance between system complexity and performance efficiency.

6.3 Future Work

Building on the groundwork laid by this thesis, several avenues for future research emerge:

- **Data Diversity**: Enhancing model robustness and generalization by expanding the diversity of training data to include more varied and challenging matting scenarios.
- **Computational Efficiency**: Optimizing the computational efficiency of our methods could broaden their applicability, particularly for real-time video matting applications.
- **Ensemble Strategies**: Investigating additional ensemble strategies and exploring different combinations of adapters and backbones could further improve matting performance.
- **Real-world Applications**: Applying and rigorously testing these models in real-world video matting tasks will provide valuable insights into their practical effectiveness and areas for improvement.

Bibliography

[1] M. McGuire, W. Matusik, H. Pfister, J. F. Hughes, and F. Durand, "Defocus video matting," *ACM Transactions on Graphics (ToG)*, vol. 24, no. 3, pp. 567–576, 2005.

[2] J. Li, V. Goel, M. Ohanyan, S. Navasardyan, Y. Wei, and H. Shi, "Vmformer: End-to-end video matting with transformer," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6678–6687, 2024.

[3] W.-L. Huang and M.-S. Lee, "End-to-end video matting with trimap propagation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14337–14347, 2023.

[4] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3523–3542, 2021.

[5] U. Ruby and V. Yendapalli, "Binary cross entropy with deep learning technique for image classification," *Int. J. Adv. Trends Comput. Sci. Eng*, vol. 9, no. 10, 2020.

[6] N. Xu, B. Price, S. Cohen, and T. Huang, "Deep image matting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2970–2979, 2017.

[7] Y.-Y. Chuang, B. Curless, D. H. Salesin, and R. Szeliski, "A bayesian approach to digital matting," in *Proceedings of the 2001 IEEE Computer Society Conference*

on Computer Vision and Pattern Recognition. CVPR 2001, vol. 2, pp. II–II, IEEE, 2001.

[8] Y.-Y. Chuang, A. Agarwala, B. Curless, D. H. Salesin, and R. Szeliski, "Video matting of complex scenes," in *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pp. 243–248, 2002.

[9] R. Vilalta and Y. Drissi, "A perspective view and survey of meta-learning," *Artificial intelligence review*, vol. 18, pp. 77–95, 2002.

[10] T. Hastie, S. Rosset, J. Zhu, and H. Zou, "Multi-class adaboost," *Statistics and its Interface*, vol. 2, no. 3, pp. 349–360, 2009.

[11] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[12] R. M. Cruz, R. Sabourin, G. D. Cavalcanti, and T. I. Ren, "Meta-des: A dynamic ensemble selection framework using meta-learning," *Pattern recognition*, vol. 48, no. 5, pp. 1925–1935, 2015.

[13] M. Cheriet, J. N. Said, and C. Y. Suen, "A recursive thresholding technique for image segmentation," *IEEE transactions on image processing*, vol. 7, no. 6, pp. 918–921, 1998.

[14] S. Pare, A. Kumar, G. K. Singh, and V. Bajaj, "Image segmentation using multilevel thresholding: a research review," *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, vol. 44, pp. 1–29, 2020.

[15] M. Brejl and M. Sonka, "Object localization and border detection criteria design in edge-based image segmentation: automated learning from examples," *IEEE Transactions on Medical imaging*, vol. 19, no. 10, pp. 973–985, 2000.

[16] M. Radhakrishnan, A. Panneerselvam, and N. Nachimuthu, "Canny edge detection model in mri image segmentation using optimized parameter tuning method," *Intell. Autom. Soft Comput*, vol. 26, no. 4, pp. 1185–1199, 2020.

[17] S. Gould, T. Gao, and D. Koller, "Region-based segmentation and object detection," *Advances in neural information processing systems*, vol. 22, 2009.

[18] L. Lalaoui and T. Mohamadi, "A comparative study of image region-based segmentation algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 6, 2013.

[19] G. B. Coleman and H. C. Andrews, "Image segmentation by clustering," *Proceedings of the IEEE*, vol. 67, no. 5, pp. 773–785, 1979.

[20] G. Ramella and G. Sanniti di Baja, "Color histogram-based image segmentation," in *International Conference on Computer Analysis of Images and Patterns*, pp. 76–83, Springer, 2011.

[21] S. Ghosh, N. Das, I. Das, and U. Maulik, "Understanding deep learning techniques for image segmentation," *ACM computing surveys (CSUR)*, vol. 52, no. 4, pp. 1–35, 2019.

[22] S. C. Yurtkulu, Y. H. Şahin, and G. Unal, "Semantic segmentation with extended deeplabv3 architecture," in *2019 27th Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, IEEE, 2019.

[23] Q. Chen, T. Ge, Y. Xu, Z. Zhang, X. Yang, and K. Gai, "Semantic human matting," in *Proceedings of the 26th ACM international conference on Multimedia*, pp. 618–626, 2018.

[24] T. Fan, G. Wang, Y. Li, and H. Wang, "Ma-net: A multi-scale attention network for liver and tumor segmentation," *IEEE Access*, vol. 8, pp. 179656–179665, 2020.

[25] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," *arXiv preprint arXiv:1805.10180*, 2018.

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

[27] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

[28] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "Unet 3+: A full-scale connected unet for medical image segmentation," in *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 1055–1059, IEEE, 2020.

[29] A. Berman, A. Dadourian, and P. Vlahos, "Method for removing from an image the background surrounding a selected object," Oct. 17 2000. US Patent 6,134,346.

[30] J. Sun, J. Jia, C.-K. Tang, and H.-Y. Shum, "Poisson matting," in *ACM SIGGRAPH 2004 Papers*, pp. 315–321, 2004.

[31] A. Levin, D. Lischinski, and Y. Weiss, "A closed-form solution to natural image matting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 2, pp. 228–242, 2007.

[32] C. Rhemann, C. Rother, J. Wang, M. Gelautz, P. Kohli, and P. Rott, "A perceptually motivated online benchmark for image matting," in *Proceddings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[33] S. Cai, X. Zhang, H. Fan, H. Huang, J. Liu, J. Liu, J. Liu, J. Wang, and J. Sun, "Disentangled image matting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8819–8828, 2019.

[34] H. Lu, Y. Dai, C. Shen, and S. Xu, "Index networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 242–255, 2020.

[35] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

[36] M. Forte and F. Pitié, "f, b, alpha matting," *arXiv preprint arXiv:2003.07711*, 2020.

[37] M. A. H. Talpur, "A modular approach to image matting," 2022-11-01.

[38] X. Shen, X. Tao, H. Gao, C. Zhou, and J. Jia, "Deep automatic portrait matting," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pp. 92–107, Springer, 2016.

[39] R. Deora, R. Sharma, and D. S. S. Raj, "Salient image matting," *arXiv preprint arXiv:2103.12337*, 2021.

[40] Y. Zhang, L. Gong, L. Fan, P. Ren, Q. Huang, H. Bao, and W. Xu, "A late fusion cnn for digital matting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7469–7478, 2019.

[41] J. Li, J. Zhang, S. J. Maybank, and D. Tao, "Bridging composite and real: Towards end-to-end deep image matting," 2021.

[42] J. Li, S. Ma, J. Zhang, and D. Tao, "Privacy-preserving portrait matting," in *Proceedings of the 29th ACM international conference on multimedia*, pp. 3501–3509, 2021.

[43] N. Apostoloff and A. Fitzgibbon, "Bayesian video matting using learnt image priors," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 1, pp. I–I, IEEE, 2004.

[44] I. Choi, M. Lee, and Y.-W. Tai, "Video matting using multi-frame nonlocal matting laplacian," in *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*, pp. 540–553, Springer, 2012.

[45] D. Li, Q. Chen, and C.-K. Tang, "Motion-aware knn laplacian for video matting," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3599–3606, 2013.

[46] S.-Y. Lee, J.-C. Yoon, and I.-K. Lee, "Temporally coherent video matting," *Graphical Models*, vol. 72, no. 3, pp. 25–33, 2010.

[47] Y. Zhang, C. Wang, M. Cui, P. Ren, X. Xie, X.-S. Hua, H. Bao, Q. Huang, and W. Xu, "Attention-guided temporally coherent video object matting," in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 5128–5137, 2021.

[48] S. Lin, L. Yang, I. Saleemi, and S. Sengupta, "Robust high-resolution video matting with temporal guidance," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 238–247, 2022.

[49] H. Wu, S. Zheng, J. Zhang, and K. Huang, "Fast end-to-end trainable guided filter," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1838–1847, 2018.

[50] Z. Ke, J. Sun, K. Li, Q. Yan, and R. W. H. Lau, "Modnet: Real-time trimap-free portrait matting via objective decomposition," 2022.

[51] Y. Sun, G. Wang, Q. Gu, C.-K. Tang, and Y.-W. Tai, "Deep video matting via spatio-temporal alignment and aggregation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6975–6984, 2021.

[52] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[53] B. Koonce and B. Koonce, "Resnet 50," *Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization*, pp. 63–72, 2021.

[54] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[55] C. Rhemann, C. Rother, J. Wang, M. Gelautz, P. Kohli, and P. Rott, "A perceptually motivated online benchmark for image matting," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1826–1833, 2009.

[56] S. Lin, A. Ryabtsev, S. Sengupta, B. L. Curless, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Real-time high-resolution background matting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8762–8771, 2021.

[57] J. Li, J. Zhang, and D. Tao, "Deep automatic natural image matting," *arXiv preprint arXiv:2107.07235*, 2021.

[58] R. Zhao, B. Qian, X. Zhang, Y. Li, R. Wei, Y. Liu, and Y. Pan, "Rethinking dice loss for medical image segmentation," in *2020 IEEE International Conference on Data Mining (ICDM)*, pp. 851–860, IEEE, 2020.