COIN-Matting: Confounder Intervention for Image Matting

Zhaohe Liao[1], Jiangtong Li[2], Jun Lan[3], Huijia Zhu[3], Weiqiang Wang[3], Li Niu[*1,4], and Liqing Zhang[*1]

[1] Shanghai Jiao Tong University
{zhaoheliao, ustcnewly, zhang-lq}@sjtu.edu.cn
[2] Tongji University
jiangtongli@tongji.edu.cn
[3] Ant Group
{yelan.lj, huijia.zhj, weiqiang.wwq}@antgroup.com
[4] miguo.ai

Abstract. Deep learning methods have significantly advanced the performance of image matting. However, dataset biases can mislead the matting models to biased behavior. In this paper, we identify the two typical biases in existing matting models, specifically **contrast bias** and **transparency bias**, and discuss their origins in matting datasets. To address these biases, we model the image matting task from the perspective of causal inference and identify the root causes of these biases: the confounders. To mitigate the effects of these confounders, we employ causal intervention through backdoor adjustment and introduce a novel *model-agnostic* cofounder intervened (COIN) matting framework. Extensive experiments across various matting methods and datasets have demonstrated that our COIN framework can significantly diminish such biases, thereby enhancing the performance of existing matting models.

Keywords: Image matting · Intervention · Dataset bias

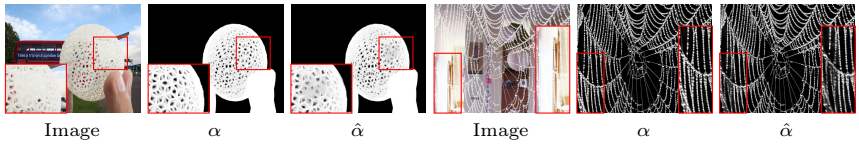
1 Introduction

Image matting aims at separating accurate foreground objects from natural images and precisely estimating the opacity on the boundary of the object. It serves as the basis for image editing, live-streaming, virtual meetings, and movie production [2, 16, 22, 30, 34]. Given an image $\boldsymbol{I}$, it is formulated as:

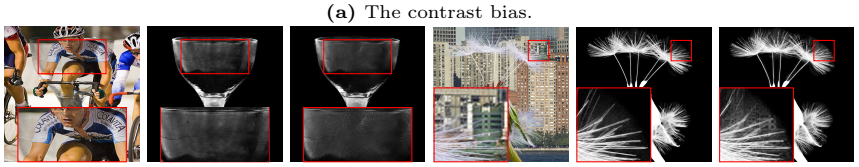
$$\boldsymbol{I} = \boldsymbol{F} \circ \boldsymbol{\alpha} + \boldsymbol{B} \circ (1 - \boldsymbol{\alpha}), \tag{1}$$

where $\boldsymbol{F}$ represents the foreground object, $\boldsymbol{B}$ denotes the background, $\boldsymbol{\alpha} \in [0, 1]$ is the opacity map, and $\circ$ denotes element-wise multiplication. As shown in Eq. (1), solving the 7 unknown values in $\{\boldsymbol{F}, \boldsymbol{B}, \boldsymbol{\alpha}\}$ from the given 3 values in $\boldsymbol{I}$ is a highly ill-posed problem. Therefore, some matting methods require

*The corresponding authors.



(a) The contrast bias.



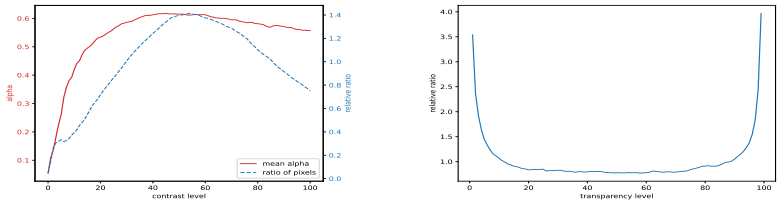
(b) The transparency bias.

Fig. 1: The observed contrast bias and transparency bias for MatteFormer [25] on the Compositional-1K dataset [40]. α and $\hat{\alpha}$ denote the ground truth and predicted alpha mattes, respectively. Images are best viewed when **zoomed in**.

additional user input as guidance, leading to a classification of matting methods based on whether and what type of user input is used. Trimap-based methods [40] are the most widely used, but obtaining a trimap can be challenging in some scenarios. Therefore, more flexible guidance, such as background [16,30] and user interaction [6, 37, 38], have been explored. Moreover, some methods [12, 38] can also work without user guidance, which is concluded as guidance-free methods.

The rapid advances in deep learning have significantly boosted the performance of image matting. However, the matting dataset may contain biases, which may further lead to the biased behavior of image matting models. We identify two primary types of bias in existing models: **contrast bias** and **transparency bias**. To illustrate these biases, we present some results of a state-of-the-art (SOTA) model, namely MatteFormer [25], in Fig. 1. The **contrast bias** refers to the correlation between the contrast and the predicted alpha value. Specifically, the model tends to predict lower alpha values in areas with low contrast between foreground and background, as highlighted by the red boxes in Fig. 1a. Regarding the **transparency bias**, it is observed that the background texture tends to emerge in the predicted alpha matte for mid-transparent foregrounds. As highlighted by the red boxes in Fig. 1b, the predicted alpha for mid-transparent area contains the background texture (*i.e.*, the texture of the rider and building for the left and right examples, respectively).

Such biased behavior of models can be explained by the biases present in the dataset. For the **contrast bias**, we calculate the alpha value at different contrast levels within the training dataset as shown in Fig. 2a. We observe that the mean alpha level in low contrast areas (*i.e.*, contrast < 20 in the CIEDE2000 color difference formula [31]) is significantly lower than that in other areas. Such dataset bias directly contributes to the observed contrast bias as the model may falsely learn to predict alpha from contrast. Additionally, the number of samples in the low contrast area is significantly fewer than in other areas, making the model more easily led by such dataset bias in these areas. Regarding the



(a) The contrast bias in the dataset. **(b)** The transparency bias in the dataset.

Fig. 2: The observe biases in matting datasets.

transparency bias, we analyze the distribution of transparency levels within the training dataset in Fig. 2b. We observe that most foreground objects are either highly salient or very opaque. Therefore, the alpha matte is expected to be either close to zero or exhibit textures in the image. For the mid-transparency area, whose alpha matte clearly differs from zero, the model is misled to focus on the texture of the image. Thus, the texture of the background in such areas may appear in the alpha matte, explaining the appearance of transparency bias.

To address these biases, we model the image matting task from the perspective of causal inference, and utilize the power of intervention to reduce the influence of the biases. The causal graph for image matting is depicted in Fig. 3a, where $\mathcal{I}, \mathcal{F}, \mathcal{B}$, and $\mathcal{A}$ are variables representing the image, foreground, background, and alpha matte respectively, while $\mathcal{C}$ and $\mathcal{T}$ are **confounder** variables denoting the contrast and the transparency of foreground. The notation $\mathcal{B} \leftarrow \mathcal{I} \rightarrow \mathcal{F}$ suggests that an image can be decomposed into foreground and background, underpinning the basic assumption of image matting. $\mathcal{F} \rightarrow \mathcal{A}$ implies that the alpha matte is directly influenced by the foreground object, indicating that changes in the background should not affect the foreground's representation. The **confounder** variables, representing the common causes affecting multiple variables in the foreground, background, and image, serve as the source of biases. Typically, these can include properties of visual objects. Given that both the foreground and background consist of objects, these properties typically cause the presentation of the foreground, background, and images are presented.

However, directly measuring all possible confounders may be impractical, as we can not enumerate all properties of visual objects. Nevertheless, based on observed dataset biases, we can identify the most typical confounders. As highlighted in Fig. 3a, we focus on two typical confounders: **contrast** $\mathcal{C}$ and **transparency** $\mathcal{T}$, which underlie the observed contrast bias and transparency bias, respectively. Contrast measures the color difference between foreground and background objects, while transparency refers to the opacity of the foreground. Therefore, we conclude $\mathcal{F} \leftarrow \mathcal{T} \rightarrow \mathcal{I}$ and $\mathcal{C} \rightarrow \mathcal{X}$ for all $\mathcal{X} \in \{\mathcal{F}, \mathcal{I}, \mathcal{B}\}$ by their definition. The dotted line indicates a potential correlation between $\mathcal{C}$ and $\mathcal{T}$, suggesting that we cannot definitively assert that $\mathcal{C} \perp\!\!\!\perp \mathcal{T}$. While existing matting methods generally focus on maximizing the likelihood $P(\mathcal{A}|\mathcal{I})$, the presence of backdoor paths $\mathcal{B} \leftarrow \mathcal{C} \rightarrow \mathcal{F} \rightarrow \mathcal{A}$, $\mathcal{B} \leftarrow \mathcal{I} \leftarrow \mathcal{C} \rightarrow \mathcal{F} \rightarrow \mathcal{A}$, and $\mathcal{B} \leftarrow \mathcal{I} \leftarrow \mathcal{T} \rightarrow$

(a) The causal graph. **(b)** The backdoor adjustment.

Fig. 3: The causal graph and causal intervention for image matting.

$\mathcal{F} \rightarrow \mathcal{A}$ results in $\mathcal{B} \not\perp\!\!\!\perp \mathcal{A}|\mathcal{I}$. This lack of conditional d-separation between $\mathcal{B}$ and $\mathcal{A}$ elucidates the co-occurrence of background textures in specific transparency regions and the contrast-induced alpha bias.

To remove the effect of these confounders, we propose our *model-agnostic* cofounder intervened (COIN) matting framework. Specifically, we explore the **do-calculus** to intervene in the influence of $\mathcal{T}$ and $\mathcal{C}$ via backdoor adjustment, as shown in Fig. 3b. The do-calculus actively assigns values to variables, rather than passively observing them from the data. Therefore, the do-calculus is an intervention on the influence of their direct causes, which is used for **deconfounding.** While the ideal way to calculate the do-operator involves collecting all images with varying contrast and transparency into the training data, this approach is evidently impractical. Furthermore, the high cost of labeling the alpha matte makes it exceedingly difficult to even approximate the aforementioned intervention "physically". To address this issue, we apply backdoor adjustment to calculate $P(\mathcal{A}|do(\mathcal{F},\mathcal{B},\mathcal{I}))$, as described in Secs. 3.2 to 3.4. By actively assigning values to $\mathcal{F},\mathcal{B}$, and $\mathcal{I}$, we block backdoor paths and magnify the influence of cofounders, thus enhancing the robustness of the matting results against the previously mentioned dataset biases. By applying our framework to state-of-the-art (SOTA) image matting methods with different types of user guidance (*i.e.*, MatteFormer [25], DIIM [6], and dugMatting [38]) and two datasets (*i.e.*, Comp1K and Distinct-646), we verified the effectiveness of our framework through extensive experiments. Our contributions are summarized as follows:

1. We examine the biases in matting methods, analyzing how dataset biases cause them and affect alpha matte prediction.
2. We introduce a novel and *model-agnostic* confounder-intervened matting framework that employs causal intervention to sever the influence of confounders, thereby mitigating the aforementioned biases.
3. Through comprehensive experiments on various matting methods and multiple datasets, we demonstrate that our framework significantly enhances the performance of SOTA image matting methods.

2 Related Work

2.1 Image Matting

As described in Sec. 1, image matting presents a highly ill-posed problem. The solutions can be classified based on the user input prior they rely upon.

Among the prior-based image matting methods, the trimap is the most widely utilized prior [1,3,4,9,14,18,21,23,23,25,33,40–42]. Since [40] introduced the first well-known deep learning-based image matting method, featuring an encoder-decoder architecture, this framework has been extensively adopted and improved by subsequent methods. Researchers have enhanced this architecture from various perspectives, including generative adversarial training [23], matting with indices [21], affinity-based methods [3, 14], context-aware matting [9], the introduction of vision transformers [1, 25, 41], prior alignment [18], expansion to ultra-high resolution via sparsity [33], and data augmentation [4, 42].

However, trimap is not always available for real-world applications. Thus, recent researchers seek for more flexible and accessible user inputs [5, 6, 8, 13, 16, 30, 37, 38, 43]. Background matting [30] uses the background image as the prior, which can be more accessible for scenes like virtual meeting and live-streaming. It is further enhanced by recursive excitation [5] and extended into real-time version [16]. The coarse mask is used instead of trimap in [43] which shows better penitential of cooperating with the segmentation models [11]. The user feedbacks, such as clicks [6], scribbles [8], their combination [37], proposal selection [38] and even natural language [13], are also informative for matting.

2.2 Causal Inference

Causal inference [27, 29] plays a pivotal role in analyzing dataset bias and mitigating counter-causal correlations. It can be categorized into two main types: **deconfounding** and **counterfactual inference**. Serving as a fundamental statistical tool, it finds widespread application across various fields, such as image segmentation [44], image classification [35], image and video question answering [10, 15, 24], dialog systems [45], and recommendation systems [36]. Typically, causal-related methods conceptualize the task within a causal graph, where nodes symbolize variables, and edges denote their causal interactions. By examining the causal relations among variables, the causal graph facilitates identifying confounders and backdoor paths, thereby enabling the control of biases.

3 Methodology

This section outlines our intervention approach to mitigating biases introduced by confounders, as discussed in Sec. 1. In Sec. 3.1, we examine the role of confounder variables within the causal graph (*i.e.*, Fig. 3), illustrating their impact on the performance of matting models through mathematical analysis. In Sec. 3.2, the concept of backdoor adjustment is introduced to counteract biases. In Sec. 3.3, we discuss the construction of space representations for each variable in the causal graph to facilitate backdoor adjustment implementation. In Sec. 3.4, the calculation of expectations, as derived in Sec. 3.2, is explained to execute the causal intervention. Finally, in Sec. 3.5, we integrate the above components into our model-agnostic COIN matting framework. Throughout this section, we denote variables by calligraphic letters (*e.g.*, $\mathcal{A}$, $\mathcal{F}$, $\mathcal{I}$), feature spaces

by boldface capital letters (*e.g.*, $\boldsymbol{A}$, $\boldsymbol{F}$, $\boldsymbol{I}$), and feature vectors by boldface lowercase letters (*e.g.*, $\boldsymbol{a}$, $\boldsymbol{f}$, $\boldsymbol{i}$). Additional introduction of background and more detailed derivations are provided in the supplementary material.

3.1 Confounders and the Backdoor Paths

In the task of image matting, confounders are variables that are the direct cause of at least two of the following: the foreground, background, and the image. As illustrated in Fig. 3a, the observational probability can be expressed as

$$P(\mathcal{A}|\mathcal{I}) = \sum_{\boldsymbol{t}} \sum_{\boldsymbol{c}} P(\mathcal{A}|\mathcal{I}, \boldsymbol{t}, \boldsymbol{c}) P(\boldsymbol{c}, \boldsymbol{t}|\mathcal{I}), \tag{2}$$

where $\boldsymbol{t}$ and $\boldsymbol{c}$ represent the split of confounder $\mathcal{C}$ and $\mathcal{T}$, respectively. Given that the observed distribution $P(\boldsymbol{c}, \boldsymbol{t}|\mathcal{I})$ significantly deviates from a uniform distribution (as depicted in Fig. 2), a specific partition of $\langle \mathcal{C}, \mathcal{T} \rangle$ predominantly influences the observed probability $P(\mathcal{A}|\mathcal{I})$ through a large $P(\boldsymbol{c}, \boldsymbol{t}|\mathcal{I})$. For instance, in the training dataset, regions with high contrast and low transparency are prevalent, leading to their dominance in the observed probability and observed biases. Specifically, in low-contrast regions, the alpha values are typically to be lower, causing the observed contrast bias. Furthermore, the model is trained to discern detailed textures in regions with opaque foregrounds and to assign nearly zero alpha values in highly transparent areas. Since these conditions consists most of the dataset, $P(\boldsymbol{c}, \boldsymbol{t}|\mathcal{I})$ with extreme values of $\boldsymbol{t}$ significantly contributes to $P(\mathcal{A}|\mathcal{I})$. Consequently, the model tends to extract highly textured patterns in medium transparency areas as in opaque foreground areas, where alpha values are apparently not as minimal as in transparent regions. Such background textures in the predicted alpha matte are concluded as transparency bias.

By adopting a causal perspective on Fig. 3a, we identify the presence of backdoor paths introduced by the confounder variables $\mathcal{C}$ and $\mathcal{T}$. In particular, the backdoor paths $\mathcal{A} \leftarrow \mathcal{F} \leftarrow \mathcal{C} \rightarrow \mathcal{B}$, $\mathcal{B} \leftarrow \mathcal{I} \leftarrow \mathcal{C} \rightarrow \mathcal{F} \rightarrow \mathcal{A}$, and $\mathcal{A} \leftarrow \mathcal{F} \leftarrow \mathcal{T} \rightarrow \mathcal{I} \rightarrow \mathcal{B}$ result in $\mathcal{A} \not\perp\!\!\!\perp \mathcal{B}|\mathcal{I}$, elucidating the influence of the background on the predicted alpha matte. More specifically, the former two paths and the latter are directly accountable for the contrast bias and transparency bias, respectively. To intervene the effects of the confounders and sever the backdoor paths they establish, we employ the **backdoor adjustment** [27, 29] as detailed in Sec. 3.2.

3.2 Causal Intervention via Backdoor Adjustment

The **do-calculus** [27, 28] signifies active assignment of values to variables, distinguishing it from passive observation within the data, which constitutes **intervention** on the causal factors of variables. For brevity, we denote $P(do(\mathcal{X} = x))$ as $P(do(\mathcal{X}))$, implying active assignment of a value to variable $\mathcal{X}$. As illustrated in Fig. 3b, to sever all incoming causal links of $\mathcal{I}, \mathcal{F}, \mathcal{B}$ (*i.e.*, nullifying the influence of $\mathcal{C}$ and $\mathcal{T}$), we focus on computing $P(\mathcal{A}|do(\mathcal{I}, \mathcal{F}, \mathcal{B}))$. Given that $\mathcal{T}$ and

$\mathcal{C}$ block all specified backdoor paths, we can derive:

$$\begin{aligned} P(\mathcal{A}|do(\mathcal{I},\mathcal{F},\mathcal{B})) &= \sum_{\boldsymbol{t}}\sum_{\boldsymbol{c}} P(\mathcal{A}|do(\mathcal{I},\mathcal{F},\mathcal{B}),\boldsymbol{t},\boldsymbol{c})P(\boldsymbol{c},\boldsymbol{t}|do(\mathcal{I},\mathcal{F},\mathcal{B})) \\ &= \sum_{\boldsymbol{t}}\sum_{\boldsymbol{c}} P(\mathcal{A}|\boldsymbol{i},\boldsymbol{f},\boldsymbol{b},\boldsymbol{t},\boldsymbol{c})P(\boldsymbol{c},\boldsymbol{t}). \end{aligned} \tag{3}$$

Moreover, based on Fig. 3b, we observe $\mathcal{A} \perp\!\!\!\perp \boldsymbol{b}|(\boldsymbol{f},\boldsymbol{i},\boldsymbol{c},\boldsymbol{t})$ and $\mathcal{A} \perp\!\!\!\perp \boldsymbol{i}|(\boldsymbol{f},\boldsymbol{c},\boldsymbol{t})$. Thus, the equation simplifies to:

$$P(\mathcal{A}|do(\mathcal{I},\mathcal{F},\mathcal{B})) = \sum_{\boldsymbol{t}}\sum_{\boldsymbol{c}} P(\mathcal{A}|\boldsymbol{f},\boldsymbol{t},\boldsymbol{c})P(\boldsymbol{t},\boldsymbol{c}) = \mathbb{E}_{\boldsymbol{t},\boldsymbol{c}}[P(\mathcal{A}|\boldsymbol{f},\boldsymbol{t},\boldsymbol{c})]. \tag{4}$$

Therefore, to predict the alpha matte with confounders intervened, we need to model the relevant spaces and calculate the expectations in Eq. (4), which are further discussed in Sec. 3.3 and Sec. 3.4, respectively.

However, to calculate the expectation in Eq. (4), obtaining the foreground representation is also necessary. Although the only input is the image (and the possible user guidance), directly inferring the foreground representation from the image representation is biased, as the confounders similarly facilitate backdoor paths to $\mathcal{F}$ as to $\mathcal{A}$. Specifically, the backdoor paths $\mathcal{F} \leftarrow \mathcal{C} \rightarrow \mathcal{B}$, $\mathcal{B} \leftarrow \mathcal{I} \leftarrow \mathcal{C} \rightarrow \mathcal{F}$, and $\mathcal{F} \leftarrow \mathcal{T} \rightarrow \mathcal{I} \rightarrow \mathcal{B}$ lead to $\mathcal{F} \not\perp\!\!\!\perp \mathcal{B}|\mathcal{I}$, indicating that the background's transparency and contrast similarly affect the foreground as influencing the prediction of the alpha matte. To address this issue, we also conduct an intervention for predicting the foreground, yielding:

$$\begin{aligned} P(\mathcal{F}|do(\mathcal{I},\mathcal{B})) &= \sum_{\boldsymbol{t}}\sum_{\boldsymbol{c}} P(\mathcal{F}|\boldsymbol{i},\boldsymbol{b},\boldsymbol{t},\boldsymbol{c})P(\boldsymbol{c},\boldsymbol{t}) \\ &= \sum_{\boldsymbol{t}}\sum_{\boldsymbol{c}} P(\mathcal{F}|\boldsymbol{i},\boldsymbol{t},\boldsymbol{c})P(\boldsymbol{c},\boldsymbol{t}) = \mathbb{E}_{\boldsymbol{c}}\mathbb{E}_{\boldsymbol{t}}[P(\mathcal{F}|\boldsymbol{i},\boldsymbol{t},\boldsymbol{c})], \end{aligned} \tag{5}$$

where the multi-scale image representation $\boldsymbol{i} = \langle \boldsymbol{i}_1, \boldsymbol{i}_2, \ldots, \boldsymbol{i}_n \rangle$ is parameterized by the matting backbone, aligned with the scales of $\boldsymbol{C}$ and $\boldsymbol{T}$ as described in Sec. 3.3. The expectations are calculated as detailed in Sec. 3.4.

3.3 Space Representations

Transparency Space $\boldsymbol{T}$. Given the importance of multi-scale information in image matting, we divide the transparency space into several sub-spaces across different scales. Specifically, the transparency space $\boldsymbol{T} = \langle \boldsymbol{T}_1, \boldsymbol{T}_2, \ldots, \boldsymbol{T}_n \rangle$, with $\boldsymbol{T}_i$ representing the transparency at the i-th scale, and n indicating the total number of sub-spaces. The i-th scale denotes the transparency of the image patches sized $2^i \times 2^i$. Considering the average alpha value as an approximation for the "transparency level" of an area, we define $\boldsymbol{T}_i$ as the average alpha value of pixels within the i-th scale patches. This formulation allows us to discretize the transparency sub-spaces further by categorizing the transparency level into K_t distinct classes, given its bounded nature within $[0, 1]$. The construction of the feature space entails the following steps:

1. Utilize a pretrained image feature extractor to derive the patch features at the i-th scale, for each $i \in \{1, \ldots, n\}$.
2. Determine the transparency level category $k \in \{1, \ldots, K_t\}$ for every patch.
3. Compute the average patch representation for all patches within the k-th category as the representation of this transparency category.

Hence, we formulate the feature space $\boldsymbol{T}_i \in \mathbb{R}^{d_i \times K_t}$, where d_i signifies the dimensionality of the feature extracted at the i-th scale.

Contrast Space $\boldsymbol{C}$. Similarly to the transparency space, the contrast space is partitioned into multiple sub-spaces across various scales. We define $\boldsymbol{C} = \langle \boldsymbol{C}_1, \boldsymbol{C}_2, \ldots, \boldsymbol{C}_n \rangle$, with each $\boldsymbol{C}_i$ representing contrast at the i-th scale, and n marking the total number of sub-spaces. Each i-th scale denotes the contrast of the foreground and background within $2^i \times 2^i$ image patches.

Quantifying contrast, however, presents more complexity than measuring transparency. Initially, we estimate the color of a patch using the average color vector of all pixels within it in the color space. Subsequently, the contrast between foreground and background patches is determined by the color distance between their average color vectors. Given the multifaceted nature of color perception, we employ the CIEDE2000 color difference formula [31] for calculating color distance, as it closely aligns with human visual perception of color differences. This formula, incorporating lightness, chroma, and hue, yields a distance within the range of $[0, 100]$. Following this approach, we categorize the contrast level into K_c discrete classes and construct the feature space $\boldsymbol{C}_i \in \mathbb{R}^{d_i \times K_c}$, mirroring the methodology used for the transparency space. The primary adjustment involves transitioning from transparency to contrast categories for the classification.

Joint Confounder Space $\boldsymbol{Z}$. Given that $\mathcal{T}$ and $\mathcal{C}$ may not be independent, we cannot decompose the joint expectation in Eq. (4) into separate expectations for each. Therefore, it becomes necessary to model the joint space of $\boldsymbol{T}$ and $\boldsymbol{C}$. An effective approach is to define the joint space as the Cartesian product of the sub-spaces. Consequently, the joint confounder space $\boldsymbol{Z} = \langle \boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_n \rangle = \langle \boldsymbol{T}_1 \times \boldsymbol{C}_1, \boldsymbol{T}_2 \times \boldsymbol{C}_2, \ldots, \boldsymbol{T}_n \times \boldsymbol{C}_n \rangle$, where each $\boldsymbol{T}_i \times \boldsymbol{C}_i \in \mathbb{R}^{d_i \times (K_t \cdot K_c)}$ represents a multidimensional space formed by the Cartesian product of transparency and contrast spaces at the i-th scale. Thus, Eqs. (4) and (5) are simplified as following:

$$\begin{aligned} P(\mathcal{A}|do(\mathcal{I}, \mathcal{F}, \mathcal{B})) &= \sum_{\boldsymbol{z}} P(\mathcal{A}|\boldsymbol{f}, \boldsymbol{z}) P(\boldsymbol{z}) = \mathbb{E}_{\boldsymbol{z}}[P(\mathcal{A}|\boldsymbol{f}, \boldsymbol{z})], \\ P(\mathcal{F}|do(\mathcal{I}, \mathcal{B})) &= \sum_{\boldsymbol{z}} P(\mathcal{F}|\boldsymbol{i}, \boldsymbol{z}) P(\boldsymbol{z}) = \mathbb{E}_{\boldsymbol{z}}[P(\mathcal{F}|\boldsymbol{i}, \boldsymbol{z})], \end{aligned} \tag{6}$$

where $\boldsymbol{z} = \langle \boldsymbol{c}, \boldsymbol{t} \rangle$ represents the split of joint observed confounders $\mathcal{Z} = \langle \mathcal{T}, \mathcal{C} \rangle$.

3.4 Expectations

The expectations are calculated over the joint confounder space $\boldsymbol{Z}$. Utilizing the Normalized Weighted Geometric Mean (NWGM) approach [32, 39], we reformulate the expectation as follows:

$$P(\mathcal{A}|do(\mathcal{I}, \mathcal{F}, \mathcal{B})) = \mathbb{E}_{\boldsymbol{z}} P(\mathcal{A}|\boldsymbol{f}, \boldsymbol{z}) = \mathbb{E}_{\boldsymbol{z}}\ \sigma(g(\boldsymbol{f}, \boldsymbol{z})) \approx \sigma(g(\boldsymbol{f}, \mathbb{E}_{\boldsymbol{z}}[\boldsymbol{z}])), \tag{7}$$

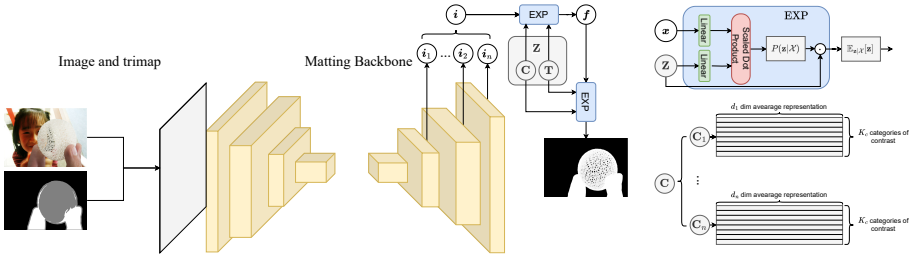


Fig. 4: The COIN framework. $\boldsymbol{i}$ and $\boldsymbol{f}$ denote the image feature and foreground feature. $\boldsymbol{T}$, $\boldsymbol{C}$ and $\boldsymbol{Z}$ are transparency space, contrast space, and their joint confounder space respectively, which are constructed in Sec. 3.3. Spaces corresponding to the confounder variable are marked gray. EXP represents the expectations described in Sec. 3.4. $\sigma(\cdot)$ and $g(\cdot)$ are ignored for simplicity. $\odot$ represents the dot production.

where $\sigma(\cdot)$ denotes a sigmoid function and $g(\cdot)$ represents a linear transformation for each sub-space. Similarly, for the foreground parameterization:

$$P(\mathcal{F}|do(\mathcal{I},\mathcal{B})) = \mathbb{E}_{\boldsymbol{z}} P(\mathcal{F}|\boldsymbol{i},\boldsymbol{z}) = \mathbb{E}_{\boldsymbol{z}}\ g(\boldsymbol{i},\boldsymbol{z}) = g(\boldsymbol{i},\mathbb{E}_{\boldsymbol{z}}[\boldsymbol{z}]). \tag{8}$$

This approach shifts the expectations to the feature level. To prevent model collapse, we use $\mathbb{E}_{\boldsymbol{z}|\mathcal{F}}$ and $\mathbb{E}_{\boldsymbol{z}|\mathcal{I}}$ as approximations for $\mathbb{E}_{\boldsymbol{z}}$ respectively. To emulate these feature-level expectations, we leverage a sub-space-wise attention mechanism. Take $\mathbb{E}_{\boldsymbol{z}|\mathcal{I}}[\boldsymbol{z}]$ as an example. As spaces $\boldsymbol{Z}_i$ are heterogeneous with the feature spaces of $\boldsymbol{i}_j$ when $i \neq j$, we apply the attention mechanism for each scale respectively. Therefore, the expectation is

$$\mathbb{E}_{\boldsymbol{z}|\mathcal{I}}[\boldsymbol{z}] = \left\langle \mathbb{E}_{\boldsymbol{z}_1|\mathcal{I}_1}[\boldsymbol{z}_1], \mathbb{E}_{\boldsymbol{z}_2|\mathcal{I}_2}[\boldsymbol{z}_2], \ldots, \mathbb{E}_{\boldsymbol{z}_n|\mathcal{I}_n}[\boldsymbol{z}_n] \right\rangle. \tag{9}$$

where n is the number of sub-spaces. For attention on image patches at the i-th scale, the input is the contrast sub-space $\boldsymbol{Z}_i \in \mathbb{R}^{d_i \times (K_t \cdot K_c)}$ and the image patch representation $\hat{\boldsymbol{i}}_i \in \mathbb{R}^{d_i}$ parameterized by the matting backbone, where d_i is the dimension of the latent representations. During such attention, the attention score emulates the distribution of $P(\boldsymbol{z}|\mathcal{I})$, thus the weighted sum on values ($i.e.$, $\boldsymbol{z}$) emulates the expectation $\mathbb{E}_{\boldsymbol{z}|\mathcal{I}}[\boldsymbol{z}]$. The mathematical formulation is given by:

$$\begin{aligned}
\boldsymbol{head}_i^k &= \text{softmax}(\hat{\boldsymbol{i}}_i^k \boldsymbol{W}_{i,1}^k (\boldsymbol{Z}_i \boldsymbol{W}_{i,2}^k)^T / \sqrt{d})(\boldsymbol{Z}_i \boldsymbol{W}_{i,3}^k) \\
\bar{z}_{\mathcal{I}_i} &= \mathbb{E}_{\boldsymbol{z}_i|\mathcal{I}_i}[\boldsymbol{z}_i] = \text{Concat}(\boldsymbol{head}_i^1, \boldsymbol{head}_i^2, \ldots, \boldsymbol{head}_i^{N_h}) \boldsymbol{W}_{i,o} \\
\bar{z}_{\mathcal{I}} &= \text{EXP}(\hat{\boldsymbol{i}}, \boldsymbol{Z}) = \mathbb{E}_{\boldsymbol{z}|\mathcal{I}}[\boldsymbol{z}] = \left\langle \bar{z}_{\mathcal{I}_1}, \bar{z}_{\mathcal{I}_2}, \ldots, \bar{z}_{\mathcal{I}_n} \right\rangle,
\end{aligned} \tag{10}$$

where $\boldsymbol{W}_{i,1}^k, \boldsymbol{W}_{i,2}^k, \boldsymbol{W}_{i,3}^k, \boldsymbol{W}_{i,o}$ are trainable parameters, N_h represents the number of heads, and Concat$(\cdot)$ is the concatenation operation. This process effectively captures the expectations across different confounder sub-spaces.

3.5 COIN Matting Framework

The overall structure of our COIN framework is shown in Fig. 4. The confounder space $\boldsymbol{T}$ is similarly constructed as $\mathbf{C}$, as described in Sec. 3.3. The mat-

ting backbone is employed to parameterize the image representation $\hat{\boldsymbol{i}}$. Rather than directly predicting $P(\mathcal{A}|\mathcal{I})$ as the matting backbone does, we engage in confounder intervention through backdoor adjustment to address the issues of contrast bias and transparency bias. The backdoor adjustment is implemented by the feature-level expectations. Specifically, we first process the joint space $\boldsymbol{Z}$ alongside the image representation $\boldsymbol{i}$ to compute $\mathbb{E}_{\boldsymbol{z}|\mathcal{I}}[\boldsymbol{z}]$ via attention. This expectation is subsequently concatenated with the image representations to yield the foreground representation. Similarly, the foreground representation is used to calculate $\mathbb{E}_{\boldsymbol{z}|\mathcal{F}}[\boldsymbol{z}]$ with joint confounder space $\boldsymbol{Z}$, and the expectations are further concatenated with the foreground representation to produce the alpha matte at various scales. Thus, we can express the overall framework as follows:

$$\begin{aligned} \hat{\boldsymbol{i}} &= \text{MattingBackbone}(\boldsymbol{I}); \quad \bar{\boldsymbol{z}}_{\mathcal{I}} = \text{EXP}(\hat{\boldsymbol{i}}, \boldsymbol{Z}); \quad \hat{\boldsymbol{f}} = g_f(\hat{\boldsymbol{i}}, \bar{\boldsymbol{z}}_{\mathcal{I}}); \\ \bar{\boldsymbol{z}}_{\mathcal{F}} &= \text{EXP}(\hat{\boldsymbol{f}}, \boldsymbol{Z}); \quad \hat{\boldsymbol{\alpha}} = P(\mathcal{A}|do(\mathcal{I}, \mathcal{F}, \mathcal{B})) = \sigma(g_\alpha(\hat{\boldsymbol{f}}, \bar{\boldsymbol{z}}_{\mathcal{F}})), \end{aligned} \tag{11}$$

where $g_f(\cdot)$ and $g_\alpha(\cdot)$ represent per-sub-space linear transformations, and EXP($\cdot$, $\cdot$) denotes the feature-level expectations described in Sec. 3.4.

During the training of this framework, we supervise the alpha matte at each scale, similar to the approach taken by the backbone. Furthermore, we supervise the attention score, which approximates the probability distribution of the confounders when calculating expectations, with the ground truth categories of contrast and transparency. It is important to note that a closer category index of contrast and transparency signifies a more similar meaning, as these are discretized from the continuous values of color distance and transparency. Therefore, we employ Gaussian label smoothing prior to the cross-entropy for supervising the attention score. Additional details regarding the framework can be found in the supplementary material.

4 Experiments

In this section, we evaluate our COIN framework on extensive image matting baselines across various datasets. Specifically, we integrate our framework with SOTA image matting methods across all three types of matting methods: trimap-based methods (*i.e.*, MatteFormer [25]), flexible guidance-based methods (*i.e.*, DIIM [6] an dugMatting [38]) and guidance-free methods (*i.e.*, dugMatting without human interaction [38]). We evaluate the improvement of our framework on Composition-1k [40], and Distinctions-646 [43] datasets. Additionally, we elaborate on the benefits of intervening on each confounder and the computational cost of implementing our framework in Sec. 4.3. The qualitative results are in Sec. 4.4.

4.1 Experiment Settings

Datasets. We conduct the experiments on Composition-1k [40] and Distinctions-646 [43] datasets following [25]. Composition-1k dataset consists of 431 and 50 foreground objects in the training set and test set respectively. These foregrounds

are composited with MS COCO [17] images for training and with VOC 2012 [7] images for testing. For testing, each foreground object is combined with 20 different images, generating a total of 1000 test images. The Distinctions-646 dataset contains 646 unique foreground objects, which are divided into 596 for training and 50 for testing. The composition process with background images follows the same procedure as in the Composition-1k dataset.

Metrics. For measuring the overall performance of our framework, we adopt the widely used image matting metrics [12, 25, 26, 43]: sum of absolute difference (SAD), mean square error (MSE), and the gradient (Grad) and connectivity (Conn). Additionally, to assess the improvement in reducing transparency and contrast biases, we measure SAD conditioned on each contrast and transparency category. The number of parameters and FLOPs are also reported to evaluate the computational complexity of implementing our framework.

Implementation Details. We set the number of sub-spaces as $n = 5$ for $\boldsymbol{I}, \boldsymbol{F}, \boldsymbol{A}, \boldsymbol{T}$, and $\boldsymbol{C}$. For transparency and contrast levels, we set both K_t and K_c to 25. The contrast and transparency space is initialized by the average multi-scale feature extracted by pretrained Swin-B [19, 20] as described in Sec. 3.3. We employ a $\mathcal{N}(0, 1.2)$ Gaussian distribution to smooth the label of attention score. More details can be found in in the supplementary material.

4.2 Main Results

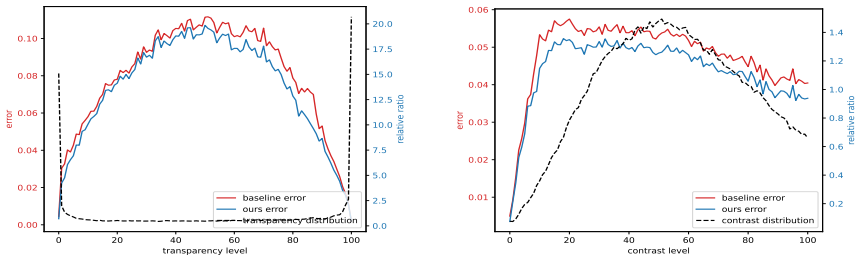
The improvement of our COIN framework is demonstrated in Tab. 1. It is evident that by intervening on the confounder variables, our COIN framework significantly enhances the performance across all matting baselines with varying guidance mechanisms on both the Composition-1k and Distinctions-646 datasets. Our framework substantially improves results for both guidance-free methods (*e.g.*, dugMatting (0-sel)) and flexible-gudance based methods (*e.g.*, dugMatting (1-sel) and DIIM). Additionally, even for the SOTA trimap-based baseline (*e.g.*, MatteFormer), our COIN framework achieves significant performance gains, decreasing SAD by 1.9 and 2.6, and MSE by 0.7 and 1.5 on the two datasets respectively. Notably, our framework tends to exhibit greater improvements on the Distinctions-646 dataset compared to the Composition-1k dataset. This discrepancy is likely attributable to the more pronounced biases present within the Distinctions-646 dataset. Furthermore, enhancements are generally more substantial for flexible guidance-based or guidance-free matting methods than for trimap-based methods, since the former methods are typically more susceptible to biases in the absence of the detailed information provided by a trimap.

To demonstrate that the enhancements are truly attributable to the reduction of contrast bias and transparency bias, we further evaluate the SAD improvements across each contrast and transparency category, as shown in Fig. 5. We selected MatteFormer as the baseline because it presents the most substantial challenge and exhibits the highest performance among the baselines. Regarding transparency, it is observable that our framework enhances the performance of baseline matting methods across all transparency levels. Furthermore, we note

Table 1: The comparison of applying our COIN framework on Composition-1k and Distinctions-646 dataset. The improvements are highlighted as superscripts.

Method	Composition-1K								Distinctions-646							
	SAD ↓		MSE ↓		Grad ↓		Conn ↓		SAD ↓		MSE ↓		Grad ↓		Conn ↓	
DIIM	37.5		10.3		18.5		34.5		36.2		11.8		19.6		35.6	
dugMatting (0-sel)*	34.1		5.7		15.6		31.2		33.2		8.8		18.2		33.2	
dugMatting (1-sel)*	25.8		4.3		9.7		22.3		24.1		7.1		12.4		24.0	
MatteFormer	23.8		4.0		8.7		18.9		21.9		6.6		11.2		20.5	
COIN(DIIM)	33.8	−3.7	8.1	−2.2	16.1	−2.4	32.0	−2.5	32.4	−3.8	9.6	−2.2	16.5	−3.1	32.3	−3.3
COIN(dugMatting (0-sel)*)	30.8	−3.3	4.5	−1.2	13.9	−1.7	28.0	−3.2	29.7	−3.5	7.4	−1.4	15.4	−2.8	30.1	−3.1
COIN(dugMatting (1-sel)*)	23.7	−2.1	3.5	−0.8	8.0	−1.7	19.1	−3.2	21.2	−2.9	6.3	−0.8	10.8	−1.6	21.1	−2.9
COIN(MatteFormer)	21.9	−1.9	3.3	−0.7	7.2	−1.5	16.8	−2.1	19.3	−2.6	5.1	−1.5	9.4	−1.8	17.8	−2.7

* k-sel denotes k times human interaction during matting.



(a) The SAD conditioned on transparency. **(b)** The SAD conditioned on contrast.

Fig. 5: The improvement of applying our framework to MatteFormer on Composition-1K dataset conditioned on each contrast and transparency category.

that the improvement is particularly pronounced in regions where the transparency level is within the range of $[50, 85]$. The training data within this range, as discussed in Sec. 1, is notably scarce but exhibits a unique alpha pattern which differs from the pattern in highly opaque or transparent foregrounds. The observed probability distribution, dominated by data from regions with high opacity or transparency, causes the transparency bias we identified. The significant enhancements in these regions indicate that our framework is reducing the transparency bias through confounder intervention. In terms of contrast, our framework not only significantly reduces the SAD error across all contrast levels compared to the baseline method but also shows that the most notable improvement occurs in areas of relatively low contrast, particularly where the contrast level ranges from $[10, 25]$. In these regions, the detail patterns increase rapidly as average contrast value and mean alpha value raises (as shown in Fig. 2a). However, owing to the insufficient data relative to other regions, these areas are most susceptible to contrast bias. Thus, the most notable improvements in low-contrast areas suggest that our framework enhances performance by alleviating the contrast bias through intervention on the contrast confounder.

Table 2: The comparison of applying intervention on different confounder variables on the Distinctions-646 dataset using MatteFormer as the baseline. The improvements are highlighted as superscripts. The GFLOPs is calculated on 1024×1024 image.

Method	Errors								Complexity	
	SAD ↓		MSE ↓		Grad ↓		Conn ↓		Parameters	FLOPs
MatteFormer	21.9		6.6		11.2		20.5		44.8M	459G
MatteFormer + $\mathcal{T}$ intervene	20.4	$^{-1.5}$	5.8	$^{-0.8}$	10.3	$^{-0.9}$	19.1	$^{-1.4}$	47.7M	468G
MatteFormer + $\mathcal{C}$ intervene	20.2	$^{-1.7}$	5.7	$^{-0.9}$	10.0	$^{-1.2}$	18.8	$^{-1.7}$	47.7M	468G
COIN(MatteFormer)	19.3	$^{-2.6}$	5.1	$^{-1.5}$	9.4	$^{-1.8}$	17.8	$^{-2.7}$	49.5M	508G

4.3 Ablation Study

Confounder Spaces. We discuss the improvement from intervening in each confounder space individually rather than their joint confounder in Tab. 2. By comparing line 3 and line 2 with line 1, it is observed that intervening on either $\mathcal{C}$ or $\mathcal{T}$ can improve the performance of the baseline methods. However, because such interventions do not block the backdoor path opened by $\mathcal{T}$ and $\mathcal{C}$ respectively, the condition $\mathcal{A} \not\perp\!\!\!\perp \mathcal{B}|\mathcal{I}$ still exists, which leads to transparency bias and contrast bias respectively, resulting in a limited performance of the intervention framework. Moreover, it was noted that the performance of intervening on $\mathcal{C}$ is slightly better than that of intervening on $\mathcal{T}$. This may be due to the fact that the contrast bias is more pronounced than the transparency bias in the dataset. By jointly intervening both $\mathcal{C}$ and $\mathcal{T}$, we may block all aforementioned backdoor paths, thus further performance improvement in the COIN framework.

Complexity Analysis. As we introduce a framework atop existing matting methods, we further analyze its space and time complexity in terms of both parameters and FLOPs to delineate the cost of integrating our framework. As indicated in Tab. 2, the additional cost imposed by our framework is minimal when compared to the baseline methods. The incremental parameters (about 10.5%) stem exclusively from the feature space and its corresponding attention mechanism, which are negligible relative to the baseline model. The only extra FLOPs (about 10.6%) originate from the computation of expectations within feature spaces, which are characterized as n attentions across $d_i \times (K_t \cdot K_c)$ matrices and $d_i \times L$ image features, where L denotes the number of image patches at the i-th scale. Such attention is also lightweight compared to the matting baseline. Overall, our framework improves the performance of the SOTA image matting baseline between 11.6% and 22.7% in different metrics with about 10% computation cost, which is remarkably efficient for integrating with the baselines.

4.4 Qualitative Results

In this section, we present qualitative comparisons of our COIN framework. In Fig. 6, we compare the alpha matting results from baseline methods with those obtained using our COIN framework. Rows 1 showcase examples illustrating the reduction of contrast bias. It is evident that the low-contrast areas

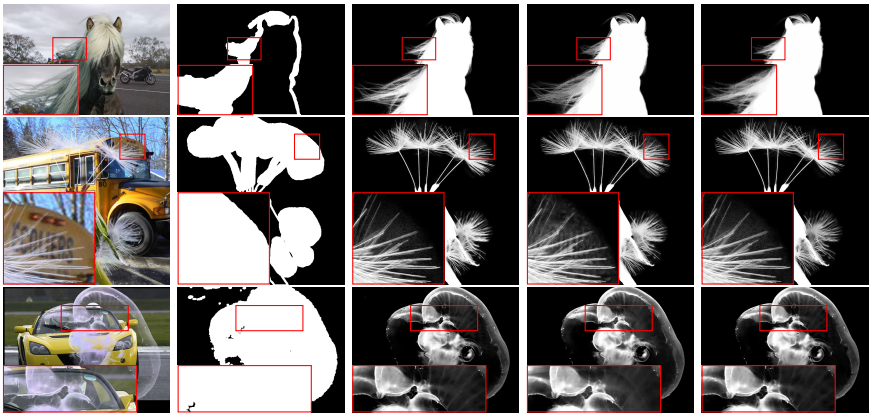


Fig. 6: The visualized comparison of applying our COIN framework on MatteFormer [25]. Images are best viewed when **zoomed in**.

highlighted in the red boxes yield alpha mattes with lower values when processed by MatteFormer. However, the integration of our COIN framework significantly diminishes this contrast bias, resulting in more precise alpha mattes. Rows 2 and 3 demonstrate the reduction of transparency bias. The highlighted mid-transparent regions in the red boxes contain background textures (*i.e.*, the texture of text and the car cage respectively) in the predicted alpha mattes. Nevertheless, with the application of our COIN framework, the alpha mattes become markedly more accurate, effectively eliminating such textures from the mid-transparent areas. Such qualitative results prove that our COIN framework can effectively reduce both contrast bias and transparency bias, therefore improving the performance of the existing image matting methods.

5 Conclusion

In this work, we have concentrated on the contrast bias and transparency bias in existing matting methods and demonstrated how dataset biases lead to biased behaviors of models. We analyzed these biases from a causal perspective and designed the confounder intervention framework to alleviate them. Extensive experiments have demonstrated that our framework markedly alleviates these biases and enhances SOTA matting methods with minor computational cost.

Acknowledgements

The work was supported by the National Natural Science Foundation of China (Grant No. 62076162), the Shanghai Municipal Science and Technology Major Project, China (Grant No. 2021SHZDZX0102) and the Postdoctoral Fellowship Program of CPSF (Grant No. GZC20241225).

References

1. Cai, H., Xue, F., Xu, L., Guo, L.: Transmatting: Enhancing transparent objects matting with transformers. In: ECCV (2022)
2. Chen, J., Li, X., Luo, L., Ma, J.: Multi-focus image fusion based on multi-scale gradients and image matting. IEEE Trans. Multim. **24**, 655–667 (2022)
3. Dai, Y., Lu, H., Shen, C.: Learning affinity-aware upsampling for deep image matting. In: CVPR (2021)
4. Dai, Y., Price, B., Zhang, H., Shen, C.: Boosting robustness of image matting with context assembling and strong data augmentation. In: CVPR (2022)
5. Deng, J., Xu, Y., Zhou, Z., He, S.: Background matting via recursive excitation. In: ICME (2022)
6. Ding, H., Zhang, H., Liu, C., Jiang, X.: Deep interactive image matting with feature propagation. IEEE Trans. Image Process. **31**, 2421–2432 (2022)
7. Everingham, M., Gool, L.V., Williams, C.K.I., Winn, J.M., Zisserman, A.: The pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. (2010)
8. Fang, X., Zhang, S., Chen, T., Wu, X., Shamir, A., Hu, S.: User-guided deep human image matting using arbitrary trimaps. IEEE Trans. Image Process. **31**, 2040–2052 (2022)
9. Hou, Q., Liu, F.: Context-aware image matting for simultaneous foreground and alpha estimation. In: ICCV (2019)
10. Li, J., Niu, L., Zhang, L.: Knowledge proxy intervention for deconfounded video question answering. In: ICCV (2023)
11. Li, J., Wang, W., Chen, J., Niu, L., Si, J., Qian, C., Zhang, L.: Video semantic segmentation via sparse temporal transformer. In: ACM MM (2021)
12. Li, J., Zhang, J., Tao, D.: Deep automatic natural image matting. In: IJCAI (2021)
13. Li, J., Zhang, J., Tao, D.: Referring image matting. In: Proceedings of the IEEE Computer Vision and Pattern Recognition (2023)
14. Li, Y., Lu, H.: Natural image matting via guided contextual attention. In: AAAI (2020)
15. Li, Y., Wang, X., Xiao, J., Chua, T.: Equivariant and invariant grounding for video question answering. In: ACM MM (2022)
16. Lin, S., Ryabtsev, A., Sengupta, S., Curless, B.L., Seitz, S.M., Kemelmacher-Shlizerman, I.: Real-time high-resolution background matting. In: CVPR (2021)
17. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: ECCV (2014)
18. Liu, Y., Xie, J., Qiao, Y., Tang, Y., Yang, X.: Prior-induced information alignment for image matting. IEEE Trans. Multim. **24**, 2727–2738 (2022)
19. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F., Guo, B.: Swin transformer v2: Scaling up capacity and resolution. In: CVPR (2022)
20. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV (2021)
21. Lu, H., Dai, Y., Shen, C., Xu, S.: Indices matter: Learning to index for deep image matting. In: ICCV (2019)
22. Lu, L., Li, J., Cao, J., Niu, L., Zhang, L.: Painterly image harmonization using diffusion model. In: ACM MM (2023)
23. Lutz, S., Amplianitis, K., Smolic, A.: Alphagan: Generative adversarial networks for natural image matting. In: BMVC (2018)

24. Niu, Y., Tang, K., Zhang, H., Lu, Z., Hua, X., Wen, J.: Counterfactual VQA: A cause-effect look at language bias. In: CVPR (2021)
25. Park, G., Son, S., Yoo, J., Kim, S., Kwak, N.: Matteformer: Transformer-based image matting via prior-tokens. In: CVPR (2022)
26. Park, K., Woo, S., Oh, S.W., Kweon, I.S., Lee, J.Y.: Mask-guided matting in the wild. In: CVPR (2023)
27. Pearl, J., Glymour, M., Jewell, N.P.: Causal inference in statistics: A primer (2016)
28. Pearl, J., Mackenzie, D.: The Book of Why: The New Science of Cause and Effect. Basic Books, Inc. (2018)
29. Rubin, D.B.: Essential concepts of causal inference: a remarkable history and an intriguing future. Biostatistics & Epidemiology (2019)
30. Sengupta, S., Jayaram, V., Curless, B., Seitz, S.M., Kemelmacher-Shlizerman, I.: Background matting: The world is your green screen. In: CVPR (2020)
31. Sharma, G., Wu, W., Dalal, E.N.: The ciede2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. Color Research and Application **30**, 21–30 (2005), `https://api.semanticscholar.org/CorpusID:29119937`
32. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. (2014)
33. Sun, Y., Tang, C., Tai, Y.: Ultrahigh resolution image/video matting with spatio-temporal sparsity. In: CVPR (2023)
34. Tan, L., Li, J., Niu, L., Zhang, L.: Deep image harmonization in dual color spaces. In: ACM MM (2023)
35. Wang, T., Zhou, C., Sun, Q., Zhang, H.: Causal attention for unbiased visual recognition. In: ICCV (2021)
36. Wang, W., Feng, F., He, X., Zhang, H., seng Chua, T.: Clicks can be cheating: Counterfactual recommendation for mitigating clickbait issue (2020)
37. Wei, T., Chen, D., Zhou, W., Liao, J., Zhao, H., Zhang, W., Hua, G., Yu, N.: Deep image matting with sparse user interactions. TPAMI (2024)
38. Wu, J., Zhang, C., Li, Z., Fu, H., Peng, X., Zhou, J.T.: dugMatting: Decomposed-uncertainty-guided matting. In: ICML (2023)
39. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A.C., Salakhutdinov, R., Zemel, R.S., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: ICML (2015)
40. Xu, N., Price, B.L., Cohen, S., Huang, T.S.: Deep image matting. In: CVPR (2017)
41. Yao, J., Wang, X., Yang, S., Wang, B.: Vitmatte: Boosting image matting with pre-trained plain vision transformers. Inf. Fusion **103**, 102091 (2024)
42. Ye, Z., Dai, Y., Hong, C., Cao, Z., Lu, H.: Infusing definiteness into randomness: Rethinking composition styles for deep image matting. In: AAAI (2023)
43. Yu, Q., Zhang, J., Zhang, H., Wang, Y., Lin, Z., Xu, N., Bai, Y., Yuille, A.L.: Mask guided matting via progressive refinement network. In: CVPR (2021)
44. Zhang, D., Zhang, H., Tang, J., Hua, X., Sun, Q.: Causal intervention for weakly-supervised semantic segmentation. In: NeurIPS (2020)
45. Zhu, Q., Zhang, W.N., Liu, T., Wang, W.Y.: Counterfactual off-policy training for neural dialogue generation. In: EMNLP (2020)