Infusing Definiteness into Randomness: Rethinking Composition Styles for Deep Image Matting

Zixuan Ye[1,†], Yutong Dai[2,†], Chaoyi Hong[1], Zhiguo Cao[1], Hao Lu[1,*]

[1] School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, China
[2] Australian Institute for Machine Learning, The University of Adelaide, Australia
hlu@hust.edu.cn

Abstract

We study the composition style in deep image matting, a notion that characterizes a data generation flow on how to exploit limited foregrounds and random backgrounds to form a training dataset. Prior art executes this flow in a completely random manner by simply going through the foreground pool or by optionally combining two foregrounds before foreground-background composition. In this work, we first show that naive foreground combination can be problematic and therefore derive an alternative formulation to reasonably combine foregrounds. Our second contribution is an observation that matting performance can benefit from a certain occurrence frequency of combined foregrounds and their associated source foregrounds during training. Inspired by this, we introduce a novel composition style that binds the source and combined foregrounds in a definite triplet. In addition, we also find that different orders of foreground combination lead to different foreground patterns, which further inspires a quadruplet-based composition style. Results under controlled experiments on four matting baselines show that our composition styles outperform existing ones and invite consistent performance improvement on both composited and real-world datasets. Code is available at: https://github.com/coconuthust/composition_styles

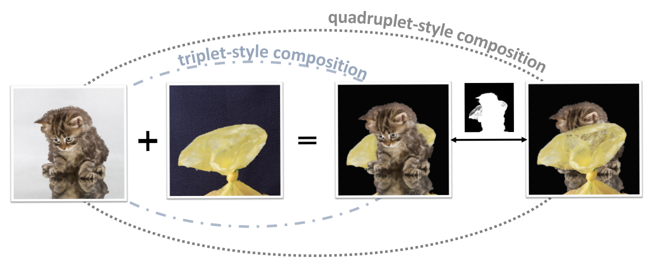

Figure 1: Two foregrounds can generate two additional combined foregrounds. Two source foregrounds and a combined one can construct a definite triplet. The combined one also has a twin combination of different foreground patterns, but shares the same alpha matte, which can be further used to form a quadruplet. They build the core of our proposed composition styles.

Introduction

Large-scale datasets provide essential ingredients for training deep neural networks (Deng et al. 2009). This sense also applies to deep image matting whose goal is to estimate the accurate alpha matte α that satisfies the matting equation

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

such that the foreground F can be separated from the background B given an image I. In particular, high-quality alpha mattes with accurate annotations are vital for training deep matting models. However, large-scale matting datasets are difficult to collect. So far only a few datasets (Xu et al. 2017; Qiao et al. 2020; Li, Zhang, and Tao 2021) with limited number of high-quality alpha mattes are publicly available. Therefore, how to maximize the value of limited alpha mattes to support the training of deep matting models has become a fundamental problem.

A common strategy is to use image composition as in Eq. (1) to synthesize samples. To form a training dataset, image composition is repetitively used following a data generation flow or a composition rule. For ease of exposition, we define such flows or rules as *composition styles*, featured by foreground selection and foreground-background composition. In the open literature, two composition styles are widely used, say DIM-style composition (Xu et al. 2017) and GCA-style composition (Li and Lu 2020). Thereinto, DIM-style composition iterates through the foreground pool; each time a chosen foreground is composited with a random background. GCA-style composition further introduces an optional process of foreground combination (Tang et al. 2019) to generate the combined foreground (Fig. 1). To distinguish from foreground-background composition, we slightly abuse the term *foreground combination* to indicate the composition between two foregrounds. Different from general-purpose augmentations that enhance appearance diversity, foreground combination boosts pattern diversity. However, we find that the naive combination of foregrounds ($\mathcal{NCF}$) used in GCA-style composition can be problematic. As shown in Fig. 2, it fails to filter out the noise and generates artifacts in the new foreground.

By probing $\mathcal{NCF}$, we find that it treats the second

[*] Corresponding author.
[†] These authors contributed equally.

foreground as background and neglects its alpha matte. To solve this, we derive a new composition equation to achieve reasonable combination of foregrounds ($\mathcal{RCF}$). In contrast to $\mathcal{NCF}$, $\mathcal{RCF}$ sequentially composites the two foregrounds onto the background. The new equation yields a different foreground representation and encodes the alpha mattes of both two foregrounds. Fig. 2 shows that $\mathcal{RCF}$ removes the artifacts in the combined foreground.

Further, we find the relation among source foregrounds and their combinations is underexplored in GCA-style composition; it executes foreground combination in a completely random manner. Such a process neglects what source foregrounds are used to form the combined foreground. Instead we consider foreground combination as a reversible chemical reaction where the source foregrounds and the combined one can mutually assist the learning of foreground patterns of each. Therefore, the occurrence frequency of a certain foreground should depend on the number of its combinations in the sample set. To meet this demand, we derive a composition style, termed *triplet-style composition*, where relevant samples are bound in a triplet to ensure the positive correlation. In this way, the training set is formed by triplet groups.

Another interesting observation is that different combination orders of two foregrounds in $\mathcal{RCF}$ can result in two distinct combined foregrounds. They share the same alpha matte but differ in foreground patterns, which forms a twin relation. We therefore realize that we can extend foreground patterns by swapping the combination order of source foregrounds. We thus introduce *quadruplet-style composition*, where a combined twin foreground is further included to form a quadruplet.

Extensive experiments under controlled conditions on four deep matting baselines, including IndexNet Matting (Lu et al. 2019), GCA Matting (Li and Lu 2020), A2U Matting (Dai, Lu, and Shen 2021) and MatteFormer (Park et al. 2022), show that our composition styles indicate a clear advantage against previous composition styles, *e.g.*, $12.7\% \sim 17.9\%$ relative improvement in the gradient metric on IndexNet Matting. Our composition styles also show a good generalization on real-world images. Moreover, we can achieve comparable performance against other composition styles even with half of the available foregrounds.

For the first time, we delve into the data generation flow in deep image matting and show that careful treatment of this flow can make a difference in matting performance.

Related Work

Deep Image Matting. Image matting is an active research area that has yielded prolific improvements during the last decades (Chen, Li, and Tang 2013; Yung-Yu Chuang et al. 2001). The emergence of deep learning further advances natural image matting. One milestone work is DIM (Xu et al. 2017), which presents an end-to-end deep matting framework and collects a benchmark dataset. Enlightened by DIM, large bodies of follow-up work emerge.

Much existing effort has been made on improving network architectures for matting. GCA (Li and Lu 2020) designs a guided contextual attention module to propagate the opacity information based on low-level features. To recover subtle details, IndexNet (Lu et al. 2019) and A2U (Dai, Lu, and Shen 2021) propose dynamic upsampling operators by predicting context-aware kernels. Recently, TIMI-Net (Liu et al. 2021) attempts to mine the information within trimap with a specialized encoder. These works lift deep matting to a new height, while they seldom probe the potential wealth outside the network architecture. Some work thinks outside the box and explores strong data augmentations. CA (Hou and Liu 2019) uses excessive data augmentations to improve the generalization of models. RMat (Dai et al. 2022) further improves data augmentations in CA to enhance the robustness of models on real-world images. In this work, we delve into the process before training begins, that is, how to form the training dataset with foreground and background pools.

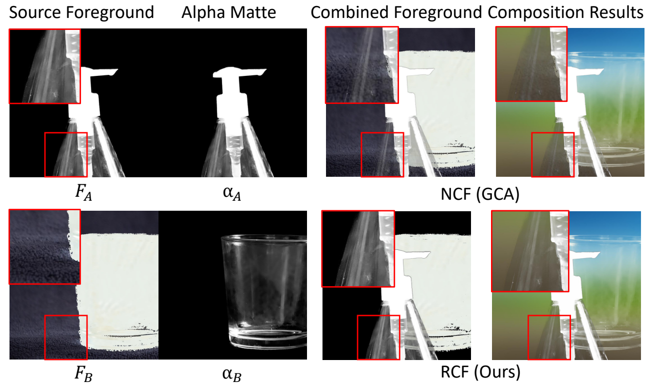

Figure 2: Combined foregrounds and composition results generated by two foreground combination operators $\mathcal{NCF}$ and our $\mathcal{RCF}$. $\mathcal{NCF}$ cannot filter out the noise in F_B, while $\mathcal{RCF}$ yields a clean composition.

Data Augmentation. Data augmentation is widely-used in deep learning to increase data diversity. It can benefit both high-level tasks such as object detection (Dwibedi, Misra, and Hebert 2017), and low-level tasks such as image enhancement (Zhang and Patel 2018) and image matting (Li and Lu 2020; Dai, Lu, and Shen 2021). They can be viewed as general-purpose techniques as no task-specific design is involved. Data augmentation is applied after the foreground selection and before the foreground-background composition. Therefore, the standard augmentations can also be add-ons to our composition styles.

In addition to standard augmentations, a stream of work adopts image composition to augment training data by leveraging its nature of generating synthetic images. For example, Mixup (Zhang et al. 2017) and CutMix (Yun et al. 2019) are adopted as data augmentations. Benefiting from the soft alpha mattes, the foreground patterns can be augmented by a combination of foregrounds based on image composition. However, we find that the treasure beneath the foreground combination remains unexplored in existing composition styles, so we delve into this research topic and particularly study the foreground selection behind foreground combination.

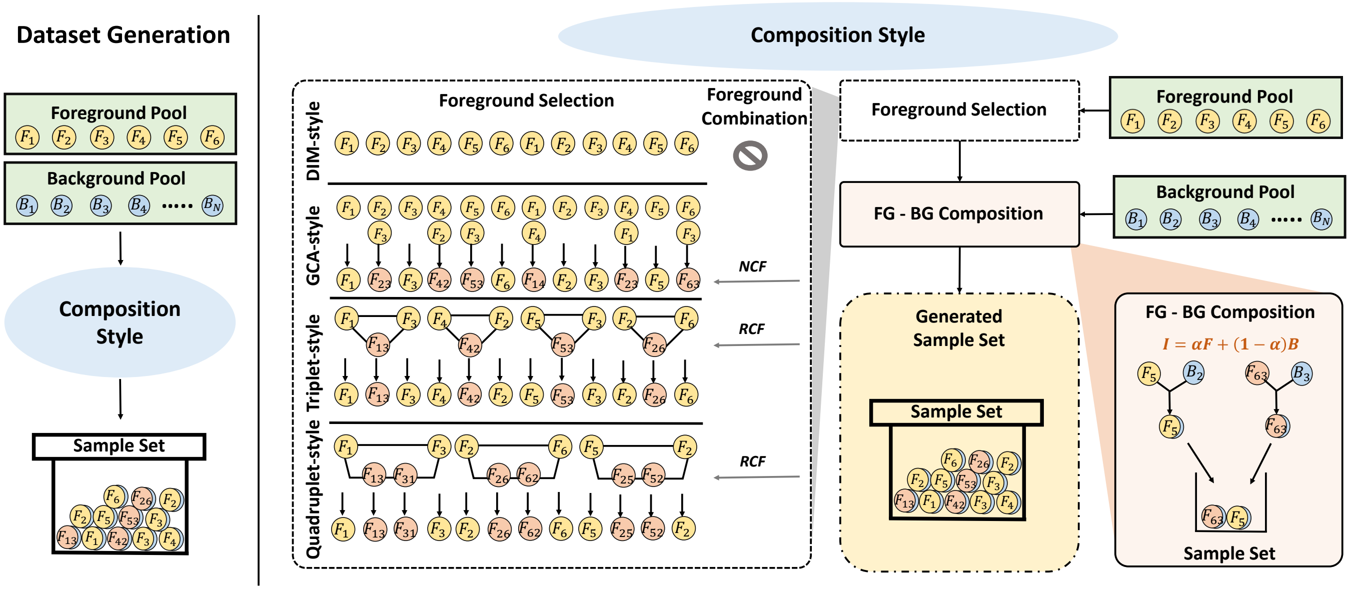

Figure 3: *Left:* Pipeline of the data generation flow. *Right:* The architecture of composition style and the comparison between previous composition styles and ours. Ours can establish a link between the source and combined foregrounds, and further increase the foreground pattern diversity brought by the source foregrounds.

A Recap of Composition Styles

Due to the lack of diverse human-annotated alpha mattes, the majority of deep matting models are trained on composited images. While previous work typically follows a data generation flow to form a training dataset with synthetic compositions, such a flow has not been paid attention to in the literature. For the first time, we formally define this flow as the composition style and use this notion to depict how to use foregrounds, backgrounds, and alpha mattes to form a (training) sample set.

Assume that we need to generate 12 samples from a foreground pool with 6 foregrounds and a background pool with infinite backgrounds. The generation process can be decomposed into two sub-stages: foreground selection and foreground-background (FG-BG) composition. As in Fig 3, foreground selection refers to how to determine the 12 foregrounds that constitute the 12 samples from the foreground pool, and FG-BG composition refers to the alpha blending (Eq. (1)) between the each foreground and a random background, which generates the sample. Each sample will be placed into a sample set after FG-BG composition. Since different composition styles share the same FG-BG composition step, the key that distinguishes different styles lies in the foreground selection.

In open literature, two composition styles are commonly used, *i.e.*, DIM-style composition (Xu et al. 2017) and GCA-style composition (Li and Lu 2020). Before we present our proposition, we first revisit these two composition styles.

DIM-Style Composition

Being the first attempt that introduces a composited training set for deep matting, DIM-style composition (Xu et al. 2017) implements naive composition by compositing each foreground onto different random backgrounds. Concretely, it iterates through the foreground pool to harvest the required number of foregrounds. For instance in Fig 3, it iterates from F_1 to F_6 twice to acquire 12 foregrounds before generating the sample set with FG-BG composition.

GCA-Style Composition

Inspired by (Tang et al. 2019), GCA-style composition (Li and Lu 2020) introduces an additional foreground combination step, where two foregrounds are combined to generate a new foreground with a probability of p. Formally, given two foregrounds F_A and F_B and their alpha mattes α_A and α_B, a new foreground F_{new} and a new alpha matte α_{new} can be generated using Eq. (1), which treats F_A as the foreground and F_B as the background. We define an operator $\mathcal{NCF}$ to characterize such naive combination of foregrounds

$$(F_{\text{new}}, \alpha_{\text{new}}) \leftarrow \mathcal{NCF}(F_A, \alpha_A, F_B, \alpha_B)\,, \tag{2}$$

where

$$F_{\text{new}} = \alpha_A F_A + (1 - \alpha_A) F_B\,, \tag{3}$$
$$\alpha_{\text{new}} = 1 - (1 - \alpha_A)(1 - \alpha_B)\,. \tag{4}$$

Akin to DIM-style composition, GCA-style composition follows the same FG-BG composition stage. They differ in foreground selection, where GCA-style combines foregrounds randomly. It first chooses source foregrounds like DIM-style composition, *i.e.*, iterating through the foreground pool. For each source foreground, it has a probability of $p = 0.5$ to be combined with another random foreground in the foreground pool. As shown in Fig 3, 6 of the 12 source foregrounds are combined with random foregrounds by $\mathcal{NCF}$.

Proposed Composition Styles

While DIM- and GCA-style compositions are effective, we show that they have certain weaknesses and may lead to sub-optimal compositions. Here we first derive a reasonable foreground combination operator $\mathcal{RCF}$ and then present our proposed composition styles built on top of $\mathcal{RCF}$.

Reasonable Combination of Foregrounds

As aforementioned, GCA-style composition blends two foregrounds F_A and F_B to generate a new foreground F_{new} and a new alpha α_{new}. We agree with the representation of α_{new}; however, we have a different opinion on F_{new}, *i.e.*, according to Eq. 3, only α_A is active, while α_B is neglected. Hence, we rethink foreground combination from a two-stage perspective and derive an alternative formulation of F_{new} by considering both α_A and α_B.

We first composite a foreground, say F_B, onto a background B to form a temporary background B_{tmp} by

$$B_{\text{tmp}} = \alpha_B F_B + (1 - \alpha_B) B \,. \tag{5}$$

At the second stage, we overlay the other foreground, say F_A, onto B_{tmp} to generate the composition I_{new} by

$$\begin{aligned} I_{\text{new}} &= \alpha_A F_A + (1 - \alpha_A) B_{\text{tmp}} \\ &= \alpha_A F_A + (1 - \alpha_A)\alpha_B F_B + (1 - \alpha_A)(1 - \alpha_B) B \,. \end{aligned} \tag{6}$$

By matching Eq. (6) with $I_{\text{new}} = \alpha_{\text{new}} F_{\text{new}} + (1 - \alpha_{\text{new}}) B$, one can infer

$$\begin{cases} \alpha_{\text{new}} F_{\text{new}} & = \alpha_A F_A + (1 - \alpha_A)\alpha_B F_B \,, \\ 1 - \alpha_{\text{new}} & = (1 - \alpha_A)(1 - \alpha_B) \,. \end{cases} \tag{7}$$

Given Eq. (7), we define a new operator $\mathcal{RCF}$ to implement reasonable combination of foregrounds such that

$$(F_{\text{new}}, \alpha_{\text{new}}) \leftarrow \mathcal{RCF}(F_A, \alpha_A, F_B, \alpha_B) \,, \tag{8}$$

where

$$F_{\text{new}} = \frac{\alpha_A F_A + (1 - \alpha_A)\alpha_B F_B}{\alpha_{\text{new}} + \epsilon} \,, \tag{9}$$

$$\alpha_{\text{new}} = 1 - (1 - \alpha_A)(1 - \alpha_B) \,. \tag{10}$$

ϵ is a small number (*e.g.*, 1e-6) used to prevent zero division. Compared with $\mathcal{NCF}$, the information of both α_A and α_B is encoded by $\mathcal{RCF}$. According to Fig. 2, α_B can play a key role in filtering out noise in F_B. Without α_B, the artifacts in F_B will passed to the combined foreground and affects the composition quality (cf. the boxed region in Fig. 2), which may also influence model learning.

Triplet-style Composition

While $\mathcal{NCF}$ and $\mathcal{RCF}$ generate the new foreground, they neglect the prior knowledge of source foregrounds. For example, from Fig. 2, the combined alpha matte exhibits new glass patterns; however, if a network does not see its source glass patterns sufficiently during training, it is likely to be confused in the inference of easy cases. To ease difficulties, we argue that the relation between the new foreground and the source foregrounds should be linked directly. Indeed, for the combined foreground, the source foregrounds can supplement the information that helps clarify the semantic meaning of complex patterns; for the source foregrounds, the inclusion of the combined foreground can assist the learning of combination rules. Therefore, by borrowing the concept of chemical reaction, we consider foreground combination as a reversible reaction

$$F_A + F_B \rightleftharpoons F_{AB} \,. \tag{11}$$

In this work, we view the source foregrounds and the combined foreground in the reaction above as *useful foregrounds*. Another important observation of this work is that, when the useful foregrounds appear with a certain proportion in the sample set, neither more nor less, the matting network can learn from foreground patterns more effectively and lead to better performance.

However, previous work (Tang et al. 2019; Li and Lu 2020) does not consider the link and neglects the complementary relation between the source and combined foregrounds. That is, the sample set can contain many useless foregrounds. To re-link the relation, we devise a novel composition style, termed triplet-style composition, which binds the useful foregrounds in a definite triplet (F_A, F_B, F_{AB}). From Fig. 3, the 12 source foregrounds construct 4 triplet groups. Samples formed by the triplet groups would have a stronger correlation than those formed by the unconstrained groups in GCA-style composition.

Quadruplet-Style Composition

The triplet-style composition builds a link between parents (source foregrounds) and a child (the combined foreground). Interestingly, we find that the child actually has a twin brother, *i.e.*, a Siamese combination. This is inspired by an observation that, if one swaps the combination order of two source foregrounds, the combinations can be different, *i.e.*,

$$F_A \text{ overlay } F_B \neq F_B \text{ overlay } F_A \,.$$

This claim can be easily validated by proof of contradiction. Assume "F_A overlay $F_B = F_B$ overlay F_A", then according to the definition of F_{new} in Eq. (7), we have

$$\frac{\alpha_A F_A + (1 - \alpha_A)\alpha_B F_B}{1 - (1 - \alpha_A)(1 - \alpha_B) + \epsilon} = \frac{\alpha_B F_B + (1 - \alpha_B)\alpha_A F_A}{1 - (1 - \alpha_A)(1 - \alpha_B) + \epsilon} \,.$$

Since $\alpha_{\text{new}} = 1 - (1 - \alpha_A)(1 - \alpha_B)$ remains unchanged, F_A and F_B must meet the following condition

$$\alpha_A \alpha_B F_B = \alpha_B \alpha_A F_A \,.$$

It is clear that our assumption is valid if and only if $F_A = F_B$. Hence, the order matters. Another point is that, while the combined twin foregrounds are different, they share the same alpha matte, just like twins who are similar in most ways but still differ in certain aspects. This can also be observed in Fig. 5. We therefore introduce the quadruplet-style composition, where an additional product has been added into the reversible reaction of the triplet-style composition (cf. Eq. (11)) as

$$F_A + F_B \rightleftharpoons F_{AB} + F_{BA} \,. \tag{12}$$

Compared with triplet-style composition, the only difference is that quadruplet-style composition generates from a foreground pair a quadruplet $(F_A, F_B, F_{AB}, F_{BA})$. From Fig. 3, the resulting foreground selection is formed by 3 quadruplet groups in quadruplet-style composition.

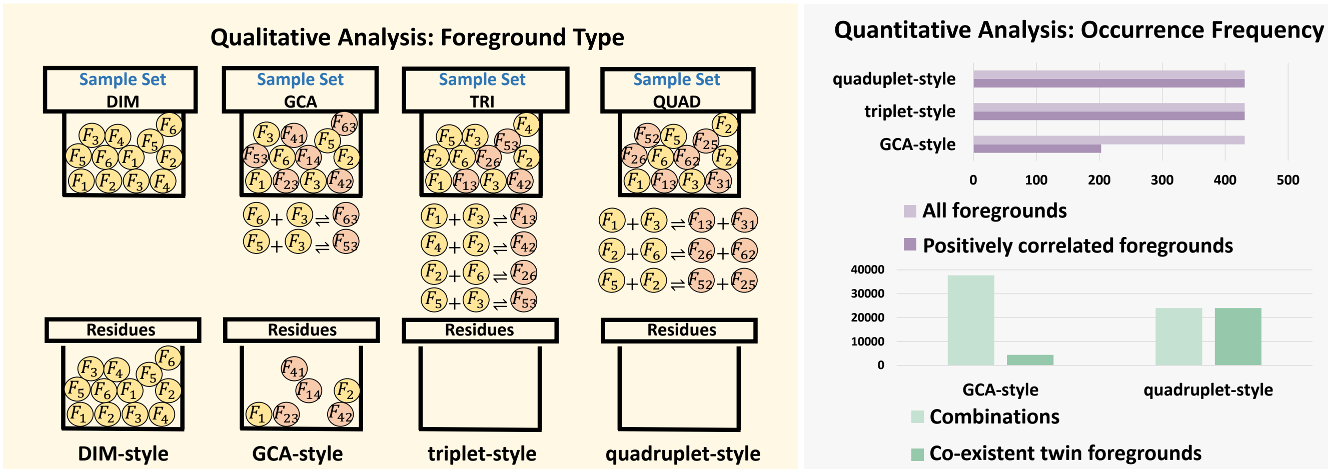

Figure 4: Foreground component analysis. *Left:* Qualitative analysis of foreground type (useful vs. useless foregrounds). After removing the useful foregrounds in the reversible reaction, the useless foregrounds are left as residues. *Right:* Quantitative analysis of occurrence frequency on i) the positively correlated foregrounds and ii) the co-existent twin foregrounds.

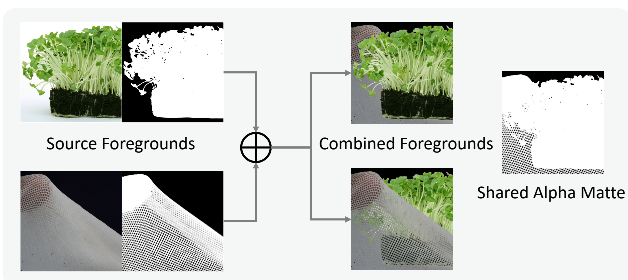

Figure 5: Combined foregrounds and alpha mattes in quadruplet-style composition.

Analysis of Different Composition Styles

DIM-style composition increases sample diversity by varying backgrounds, but does not extend foreground patterns. GCA-style composition further expands foreground diversity by foreground combination. They, however, fail to i) model the relation between the combined and source foregrounds and ii) mine the treasure beneath the two source foregrounds. We address the issues above via two novel composition styles. In particular, triplet-style composition establishes the link between three relevant foregrounds, while GCA-style composition does not have such an explicit link. To show their differences, we present *foreground component analysis*. We analyze the foreground component from two aspects: qualitative analysis on foreground type and quantitative analysis on occurrence frequency. The former distinguishes between useful and useless foregrounds based on the contribution of a certain foreground to the training of the whole sample set. The latter provides an empirical measure. Note that we ignore the influence of the background, because i) the background pool can have infinite backgrounds and ii) matting models mainly learn foreground patterns.

As shown in Fig. 4, four sample sets generated by four composition styles have different foreground components. For example, the sample set generated by DIM-style composition has no combined foreground. If removing foregrounds involved in reversible reactions, the remaining residues are the useless foregrounds. Compared with the useful ones which assist mutual learning of foreground patterns, the useless ones are only limited to learning of their own patterns. The empty residue of the sample sets generated by proposed composition styles indicate that all foregrounds are useful and contribute to effective learning of other foreground patterns.

In addition to the residues, we can also understand from the occurrence frequency of foreground. The occurrence frequency N of a single foreground should relate to that of the combinations it relates to. Formally, they should be positively correlated, *i.e.*,

$$N(F_A) \propto N(F_{AX} + F_{XA}), \quad (13)$$

where F_X is another foreground that is combined with F_A. As shown in Fig. 4, we count how many of the foregrounds follows the positive correlation; the results show that only about half of the foregrounds in GCA-style composition meet the requirement. However, our triplet-style composition allows all the foregrounds to have positive correlations by ensuring

$$N(F_A) = N(F_{AX} + F_{XA}), \quad (14)$$

while our quadruplet-style composition by

$$2N(F_A) = N(F_{AX} + F_{XA}). \quad (15)$$

To highlight the difference between quadruplet-style composition and GCA-style composition, we also count the occurrence frequency of co-existence of twin foregrounds. The statistics in Fig. 4 show that only 11.6% of the combinations in GCA-style composition own the opportunity to have their twin foregrounds, while our

	IndexNet Matting				GCA Matting				A2U Matting				MatteFormer			
	SAD	MSE	Grad	Conn	SAD	MSE	Grad	Conn	SAD	MSE	Grad	Conn	SAD	MSE	Grad	Conn
Reported	45.80	0.0130	25.90	43.70	35.28	0.0091	16.92	32.53	32.10	0.0078	16.33	29.00	23.80	0.0041	8.72	18.89
DIM-style	43.03	0.0115	22.21	41.70	37.74	0.0100	18.64	32.64	36.16	0.0089	17.20	34.06	29.29	0.0063	10.85	24.77
GCA-style	45.19	0.0130	23.62	44.68	35.16	0.0090	17.69	32.61	32.15	0.0082	16.39	29.25	23.71	0.0039	8.57	20.47
Triplet-style	38.65	0.0100	20.29	36.78	**31.91**	**0.0075**	**14.14**	**27.77**	31.52	0.0072	15.89	28.47	23.26	**0.0038**	**7.34**	19.52
Quadruplet-style	**38.16**	**0.0099**	**19.37**	**36.31**	33.16	0.0080	15.09	29.35	**30.47**	**0.0071**	**15.30**	**27.47**	**22.60**	**0.0038**	7.67	**19.26**

Table 1: Quantitative comparisons of composition styles across 4 baselines on the Composition-1K dataset. Best results are in boldface, and second-best ones are underlined. The 'Reported' row cites the published results for reference. The original IndexNet use the DIM-style composition, and the others use the GCA-style composition.

quadruplet-style composition boosts such a proportion to 1. Therefore, the quadruplet-style composition ensures that, once a combined foreground appears, its twin foreground will definitely appear in the sample set.

Results and Discussion

To verify the effectiveness of the proposed two composition styles and the foreground combination operator $\mathcal{RCF}$, we conduct extensive experiments over different matting baselines on several benchmark datasets. We evaluate on both the synthetic and real-world scenarios to validate the generalization capability.

Implementation Details

Our implementation is based on `PyTorch`. We only need to modify the dataloader without any architecture change when applying the composition styles.

Datasets. We train models on the synthetic Adobe Image Matting (Xu et al. 2017) dataset and report performance on both the synthetic Composition-1K (Xu et al. 2017) dataset and the real-world AIM-500 (Li, Zhang, and Tao 2021) and PPM-100 (Ke et al. 2022) datasets.

Evaluation Metrics. Following (Rhemann et al. 2009), we employ four standard metrics to evaluate the predicted alpha mattes, including sum of absolute differences (SAD), mean squared errors (MSE), gradient errors (Grad), and connectivity errors (Conn).

Baselines. We evaluate on four state-of-the-art matting baselines: IndexNet Matting (Lu et al. 2019), GCA Matting (Li and Lu 2020), A2U Matting (Dai, Lu, and Shen 2021), and MatteFormer (Park et al. 2022).

Fairness. All experiments are conducted under the following controlled settings to fairly compare different composition styles. First, we apply the same data augmentation strategies used in GCA Matting for different composition styles, and in this way the composition style is the only variable. Second, to eliminate the performance differences caused by different data volumes, we provide each composition style with exactly the same amount of training samples. Furthermore, we fix the order and the visual contents of the sample set when comparing the two foreground combination operators $\mathcal{NCF}$ and $\mathcal{RCF}$, to ensure that the only difference is the generation manner of F_{new}.

	AIM-500				PPM-100			
	SAD	MSE	Grad	Conn	SAD	MSE	Grad	Conn
DIM-style	27.28	0.0305	21.37	15.08	44.58	0.0264	41.83	35.35
GCA-style	28.39	0.0303	22.47	14.71	41.01	0.0249	41.15	31.87
Tri-style	**26.33**	**0.0278**	**19.61**	**12.90**	**38.94**	0.0234	**38.64**	30.21
Quad-style	26.65	0.0307	20.79	14.37	40.22	**0.0226**	40.19	**30.02**

Table 2: Quantitative comparisons of composition styles on real-world datasets with MatteFormer as the baseline.

Comparison of Composition Styles

We compare the composition styles from three aspects: effectiveness on the composited dataset, generalization ability on real-world datasets, and robustness to the number of foregrounds.

Effectiveness. We first compare the triplet- and quadruplet-style composition with prior composition styles on Composition-1K. From Table 1, both triplet- and quadruplet-style compositions consistently outperform the DIM-style and the GCA-style compositions on all baselines, suggesting that our composition styles are generic designs. Establishing a definite link reveals the value of the foreground combination, with a significant 6.54 SAD reduction on the IndexNet Matting baseline. In the open literature, such improvements often come from increased network complexity or an elaborate training strategy. However, we achieve this with only an improved data generation flow prior to model training, without any change of network design or complicated training procedures. Visual results are shown in Fig. 6. Compared with the previous styles, our proposed ones can recover more clear, detailed structure such as complex wire patterns.

Generalization. Since the composition styles are developed for composited data, it is important to know the generalization of the styles on real-world data. Here MatteFormer is chosen as the baseline, and we report performance on two real-world datasets in Table 2. Results show that models trained with our composition styles also outperform other composition styles on read-world data.

Robustness. We also explore the robustness of different styles with reduced amount of available foregrounds. In Fig. 7, the triplet-style is more robust than other styles when

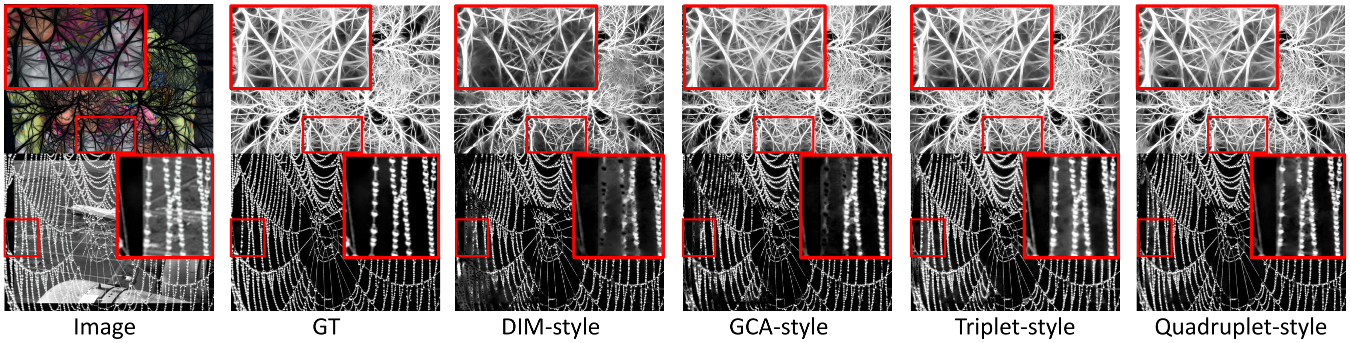

Figure 6: Qualitative results of different composition styles on Composition-1K dataset. Due to the infused definiteness in our composition styles, ours produces better visualizations on complex textures.

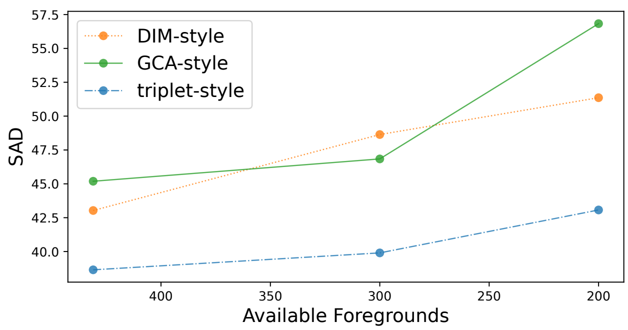

Figure 7: Robustness to the amount of available foregrounds. IndexNet is used as the baseline.

the available foregrounds decrease. Despite the foregrounds are halved, we still achieve comparable performance with the DIM-style that uses double amount of foregrounds.

Comparison of Foreground Combination

Here we compare $\mathcal{NCF}$ with the proposed $\mathcal{RCF}$ on A2U with the quadruplet-style and on IndexNet with GCA-style as the baselines, to show that the operator is independent of specific composition styles and matting models. We report performance on Composition-1K in Table 3. Results show that $\mathcal{RCF}$ outperforms $\mathcal{NCF}$ on both baselines, confirming the effectiveness and the fundamental role of $\mathcal{RCF}$.

Ablation Study of Data Preprocessing Components

We use IndexNet without any data processing as the baseline (D1). To validate the effectiveness of each component, the ablation study incorporates common data augmentation, the proposed $\mathcal{RCF}$, and the triplet-/quadruplet-style composition. Results are reported in Table 4. By comparing D1 with D2, one can see an incremental improvement of 1.84 SAD achieved by varying the visual appearance. By comparing D2 with D3 and D4, the proposed composition styles yield a substantial performance improvement. Such results corroborate the effectiveness of our composition styles and the importance of a balanced foreground distribution. In addition, there is also a marginal improvement when the triplet style is altered to the quadruplet style (D3 vs. D4).

	$\mathcal{NCF}$	$\mathcal{RCF}$	Compostion-1k SAD	MSE	Grad	Conn
A2U	✓		32.70	**0.0066**	16.24	30.29
A2U		✓	**30.47**	0.0071	**15.30**	**27.27**
IndexNet	✓		45.19	**0.0130**	**23.62**	44.68
IndexNet		✓	**44.03**	0.0135	24.76	**42.89**

Table 3: Comparison of foreground combination operator on the Composition-1K. A2U and IndexNet are adopted as baselines, respectively.

	DA	$\mathcal{RCF}$	Tri	Quad	SAD	MSE	Grad	Conn
D1					44.87	0.0124	25.08	41.23
D2	✓				43.03	0.0115	22.21	41.70
D3	✓	✓	✓		38.65	0.0100	20.29	36.78
D4	✓	✓		✓	38.16	0.0099	19.37	36.31

Table 4: Ablation study of different components. 'DA' denotes the data augmentation such as random jittering used in GCA Matting. 'Tri' and 'Quad' denote triplet- and quadruplet-style composition, respectively.

Conclusion

In this work, we introduce the concept of the composition style to characterize the data generation flow in deep image matting. We first present an improved operator $\mathcal{RCF}$ to reasonably combine foregrounds with reduced artifacts. We then propose a triplet-style composition and a quadruplet-style composition, which infuses definiteness into the random data generation flow. We validate our propositions over four state-of-the-art matting baselines on both composited and real-world datasets. Results show that our composition styles consistently outperform the previous ones and demonstrate good generalization on real scenes.

Being the first work that delves into the data generation flow in deep image matting, we believe our work points a good direction for addressing deep matting with improved composition styles.

Acknowledgement. This work is supported by the National Natural Science Foundation of China under Grant No. 62106080.

References

Chen, Q.; Li, D.; and Tang, C.-K. 2013. KNN Matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2175–2188.

Dai, Y.; Lu, H.; and Shen, C. 2021. Learning Affinity-Aware Upsampling for Deep Image Matting. In *Proceedings of IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, 6837–6846.

Dai, Y.; Price, B.; Zhang, H.; and Shen, C. 2022. Boosting Robustness of Image Matting with Context Assembling and Strong Data Augmentation. In *Proceedings of IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, 11707–11716.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, 248–255.

Dwibedi, D.; Misra, I.; and Hebert, M. 2017. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 1301–1310.

Hou, Q.; and Liu, F. 2019. Context-aware image matting for simultaneous foreground and alpha estimation. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 4130–4139.

Ke, Z.; Sun, J.; Li, K.; Yan, Q.; and Lau, R. W. 2022. MODNet: real-time trimap-free portrait matting via objective decomposition. In *Proceedings of AAAI Conference on Artificial Intelligence*, 1140–1147.

Li, J.; Zhang, J.; and Tao, D. 2021. Deep Automatic Natural Image Matting. In *Proceedings of International Joint Conference on Artificial Intelligence*, 800–806. ISBN 978-0-9992411-9-6.

Li, Y.; and Lu, H. 2020. Natural image matting via guided contextual attention. In *Proceedings of AAAI Conference on Artificial Intelligence*, 11450–11457.

Liu, Y.; Xie, J.; Shi, X.; Qiao, Y.; Huang, Y.; Tang, Y.; and Yang, X. 2021. Tripartite information mining and integration for image matting. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 7555–7564.

Lu, H.; Dai, Y.; Shen, C.; and Xu, S. 2019. Indices Matter: Learning to Index for Deep Image Matting. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 3265–3274.

Park, G.; Son, S.; Yoo, J.; Kim, S.; and Kwak, N. 2022. MatteFormer: Transformer-Based Image Matting via Prior-Tokens. In *Proceedings of IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, 11696–11706.

Qiao, Y.; Liu, Y.; Yang, X.; Zhou, D.; Xu, M.; Zhang, Q.; and Wei, X. 2020. Attention-guided hierarchical structure aggregation for image matting. In *Proceedings of IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, 13676–13685.

Rhemann, C.; Rother, C.; Wang, J.; Gelautz, M.; Kohli, P.; and Rott, P. 2009. A perceptually motivated online benchmark for image matting. In *Proceedings of IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, 1826–1833. ISBN 978-1-4244-3992-8.

Tang, J.; Aksoy, Y.; Oztireli, C.; Gross, M.; and Aydin, T. O. 2019. Learning-Based Sampling for Natural Image Matting. In *Proceedings of IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, 3050–3058. ISBN 978-1-72813-293-8.

Xu, N.; Price, B.; Cohen, S.; and Huang, T. 2017. Deep Image Matting. In *Proceedings of IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, 311–320. ISBN 978-1-5386-0457-1.

Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 6023–6032.

Yung-Yu Chuang; Curless, B.; Salesin, D.; and Szeliski, R. 2001. A Bayesian approach to digital matting. In *Proceedings of IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, II–264–II–271. ISBN 978-0-7695-1272-3.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhang, H.; and Patel, V. M. 2018. Density-aware single image de-raining using a multi-stream dense network. In *Proceedings of IEEE Conference on Computer Vision Pattern Recognition (CVPR)*, 695–704.