Disentangled Pre-training for Image Matting

Yanda Li[1], Zilong Huang[2], Gang Yu[2], Ling Chen[1], Yunchao Wei[3], Jianbo Jiao[4]
[1]University of Technology Sydney, Australia [2]Tencent
[3]Beijing Jiaotong University, China [4]University of Birmingham, UK
{liyanda95,wychao1987}@gmail.com, {zilonghuang, skicyyu}@tencent.com
Ling.Chen@uts.edu.au, j.jiao@bham.ac.uk

Abstract

Image matting requires high-quality pixel-level human annotations to support the training of a deep model in recent literature. Whereas such annotation is costly and hard to scale, significantly holding back the development of the research. In this work, we make the first attempt towards addressing this problem, by proposing a self-supervised pre-training approach that can leverage infinite numbers of data to boost the matting performance. The pre-training task is designed in a similar manner as image matting, where random trimap and alpha matte are generated to achieve an image disentanglement objective. The pre-trained model is then used as an initialisation of the downstream matting task for fine-tuning. Extensive experimental evaluations show that the proposed approach outperforms both the state-of-the-art matting methods and other alternative self-supervised initialisation approaches by a large margin. We also show the robustness of the proposed approach over different backbone architectures. Our project page is available at https://crystraldo.github.io/dpt_mat/.

1. Introduction

Image matting has played a predominant role in daily applications in the past few years, *e.g.* online meetings and smartphone applications, referring to extracting the foreground from natural images by predicting an alpha matte. A natural image $\mathcal{I}$ can be represented as a linear fusion of foreground $\mathcal{F}$ and background $\mathcal{B}$ with a weighting parameter α as defined below:

$$\mathcal{I} = \alpha\mathcal{F} + (1-\alpha)\mathcal{B}. \quad (1)$$

Matting, as a low-level computer vision problem, has been studied for decades in the literature and remains a challenging problem. With the availability of computational resources and network capacity, most existing work addresses this problem based on deep neural networks. Among these methods, the contributions mainly lie on adjusting the objective function [12,17,27], network structure [29,30,37,43] or the data input to the network [1, 33, 35]. Due to the data-driven nature of deep learning, the quality of data and its annotation is the key to the model performance, apart from the computational resources, and is becoming the bottleneck. However, due to the nature of the image matting task and its high-precision requirement for the pixel-level annotation of the alpha matte, data labelling is incredibly costly and hard to scale, restricting the development of the field significantly. Composition-1k [43] and Distinct-646 [30], the most commonly used matting datasets, only contain hundreds of foreground images with corresponding annotations, and serve as the main benchmarks for the whole community. To address this vital and timely challenge, instead of seeking more labour to annotate more data, we step back and are more interested in the question: *is that possible to leverage the freely available huge amounts of unlabelled natural images to boost the performance of image matting?*

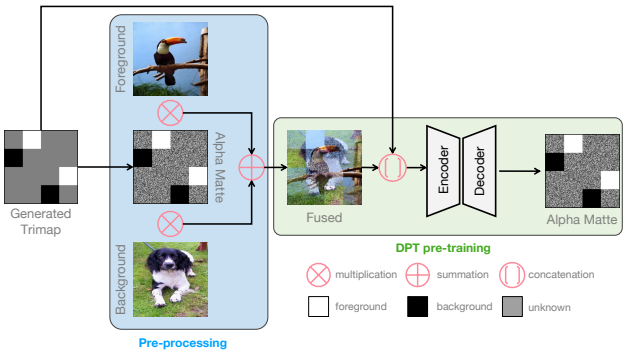

Figure 1. An overview of the proposed DPT. During pre-processing, a trimap is generated by assigning regions of foreground (white blocks), background (black blocks) and unknown (the rest grey ones). Based on that, a random alpha matte α is generated. The foreground and background images are composited with the α according to Eq. 1. Together with the trimap as input, the network is trained to estimate an alpha matte. The aforementioned generated α is used as a pseudo label to train the model.

To this end, in this work we make the first attempt towards self-supervised pre-training for image matting.

Recently, self-supervised learning (SSL) has made prominent progress in deep learning, which benefits from the rapid update of hardware. In natural language processing, self-supervised learning was first proposed to cater for the growth of millions of data. Large-scale language modelling without human-annotated labels has obtained brilliant success by autoregressive language modeling [31] and masked language modeling [10]. The idea of self-supervised learning has also greatly benefited computer vision. Lots of methods [14, 15, 41] are built to enhance the representation learning of models and achieve superior results in downstream tasks. SimCLR [6] and MoCo [15] made the breakthrough in self-supervised learning on high-level image classification representation. Several following-up works [4, 41] enhanced the ability of representation learning by improving the contrastive loss. Although visual self-supervised representation learning has proved its effectiveness in several high-level tasks, low-level tasks have barely been touched and a suitable self-supervised pre-training approach for image matting is under-explored. To bridge the gap, two challenges need to be addressed:

i) Auxiliary input is required for the matting task, *i.e.* trimap. According to Eq. 1, there are 3 unknown parameters. To estimate α given only $\mathcal{I}$ is a massively ill-posed problem, hence additional guidance is required. Existing SSL methods either learn by predicting the missing parts or constructing positive-negative pairs for contrastive learning, without considering additional guidance during the pre-training, making it infeasible to directly apply off-the-shelf SSL methods for the matting task.

ii) Image matting, by definition, is a pixel-level disentanglement problem. It not only requires foreground-background segmentation, but also needs to estimate the transparency of the fused regions near boundaries. In addition, it is a class-agnostic task, different from existing SSL methods focusing on semantic instance discrimination.

To address the above challenges, in this paper, we present a new disentangled self-supervised pre-training approach tailored for image matting, termed *DPT*. DPT is designed to be a pretext task similar to the matting task, enabling large-scale pre-training for matting-aware representation learning. In our DPT, we simulate the matting process including input guidance and supervision information on synthetic datasets which is similar to matting datasets. With the proposed DPT pre-training, the learned representation is forced to be with the potential ability of image disentanglement. Such representations are leveraged to boost the performance of image matting in the following fine-tuning stage. An illustration of the proposed DPT is shown in Figure 1. First, trimap $\boldsymbol{T}$ is generated by randomly cropping patches and splitting these patches into three categories of *background*, *foreground* and *unknown*. According to the pre-defined region information, alpha matte α is randomly generated as the pseudo label. A randomly chosen foreground image $\mathcal{F}$, and background $\mathcal{B}$ are fused according to α. The fused image is then fed into the network together with the trimap $\boldsymbol{T}$ to predict the aforementioned alpha matte. Some qualitative examples of the proposed pre-training task are shown in Figure 2.

With the proposed DPT pre-training, we show the potential of leveraging large-scale (infinite) unlabeled data for image matting with demonstrated clear improvement over the state-of-the-art. The main contributions of this work are summarised as below:

- We propose, to our knowledge, the first self-supervised large-scale pretraining approach for image matting. The pretext task is designed and tailored for the matting task and shown to be effective in learning disentanglement representations.
- The proposed DPT pre-training approach is shown to be effective across different network backbones, with consistent performance improvement.
- Extensive experimental analysis on several public datasets shows the effectiveness of the proposed method, outperforming existing image matting approaches by a large margin.

2. Related works

Natural image matting Traditional natural image matting can be summarized into two categories, sampling-based and propagation-based. Sampling-based methods [7, 13, 34] collected foreground and background colour samples to generate alpha of unknown region. Propagation-based methods [5, 18, 19] used neighbouring pixels and estimated the alpha matte of the unknown region by propagating the alpha from the foreground and background regions.

Until recently, common matting approaches are divided into two parts: trimap-based and trimap-free. Most methods [17, 26, 43] took trimap as input to provide regional information. DIM [43] introduced the most popular matting dataset Composition-1k and a two-stage encoder-decoder network with trimap as an additional input. Alphagan [27] presented a generative adversarial network for matting to predict alpha. SampleNet [37] applied sampling methods to the matting task. A novel end-to-end natural image matting method GCA [21] was proposed with a guided contextual attention module. External semantic information was incorporated into the model in SIM [36] to obtain a better alpha matte. As the first transformer-based matting model, Matteformer [29] introduced a prior token that participates in the self-attention mechanism. TransMatting [3] modelled transparent objects with a big receptive field.

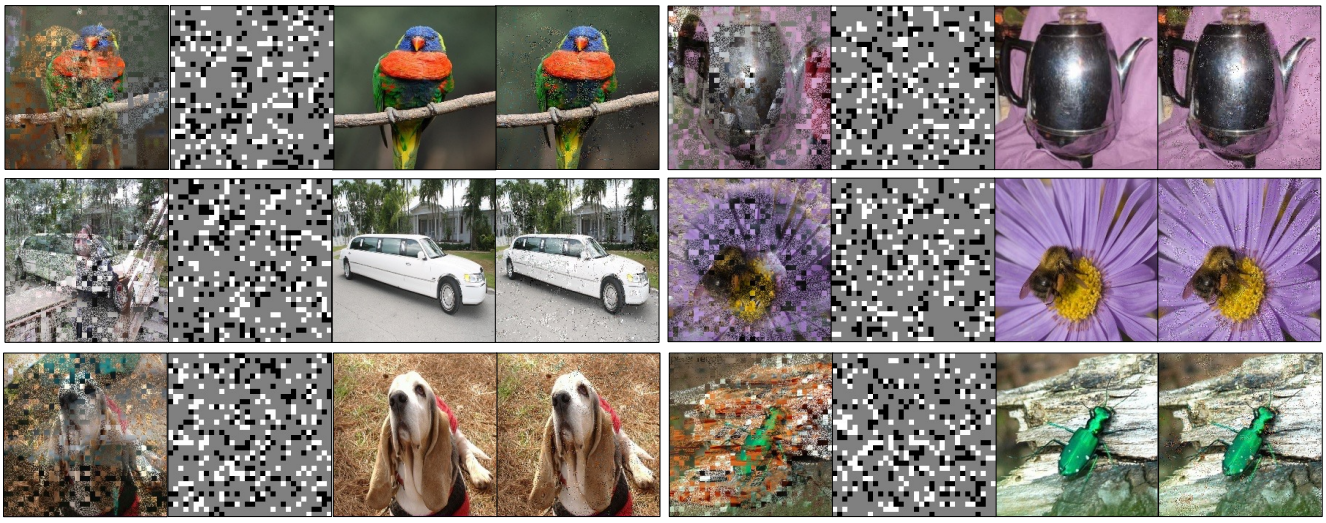

Figure 2. Example results on ImageNet-1k. For each set of images, we show the merged image (left), defined trimap (second), labelled foreground image (third) and generated foreground image (right). The size of patches in trimap is $7\times$ 7 and the unknown region occupies 75% of the patches, 12.5% for foreground and 12.5% for background. According to Eq. 1, we combine the merged image according to the labelled alpha matte and pseudo label, respectively, to generate labelled foreground and generated foreground. *Following generic trimap-based methods, the model only predicts the unknown region, and both the foreground region and background region are copied directly from the foreground image.

For trimap-free methods [20, 33], HAttMatting [30] proposed Distinctions-646 dataset and designed a hierarchical attention structure. Liu et al. [23] employed coarse annotated data coupled with fine annotated data for semantic human matting without trimaps as an extra input. Background Matting [22, 33] replaced trimap with an additional background image without the object. MG Matting [45] took coarse mask as its guidance. To control the matting with natural language description, RIM [20] proposed a new task named referring image matting to extract object alpha that matched the given language description. While there have been some outstanding trimap-free methods, however, there is a certain gap between trimap-free methods and trimap-based ones in performance. Considering trimap is widely used for image matting, we adopt trimap as the auxiliary input of the disentangled pre-training task.

Self-supervised learning Self-supervised learning approaches have been significantly used in deep learning. In natural language processing, self-supervised learning is mainly achieved by auto-regressive language modeling [31, 32] and masked language modeling [10]. These methods filled the masked portion of the input sequence by predicting the missing part for pre-training. In this way, the pre-trained model could fit hundreds of millions of data generalize well and achieve better performance when fine-tuning downstream tasks.

Motivated by the success of unsupervised learning in NLP, some self-supervised learning methods [2, 6, 14, 15, 28, 38, 40, 41] are introduced for vision tasks. MoCo [15] presented an unsupervised pre-training and transferred it to various downstream tasks by fine-tuning. MoBY [41] transferred adaption to detection and segmentation with the Swin transformer network. Masked autoencoder (MAE) [14] randomly masked patches and reconstructed the missing region. A similar idea was also verified in SimMIM [42], where the input image was masked with moderately large patch size and gained a better representation. To better adapt for dense prediction task, [28, 39] migrated contrast loss from image-level to pixel-level. CP^2 [38] facilitated both image-level and pixel-level representation through a simple copy-paste operation. There are many excellent SSL methods, whereas self-supervised learning designed for image disentanglement is under-explored.

3. Method

Our DPT is a simple self-supervised pre-training approach for image matting task. The architecture of our model is an encoder-decoder framework. We pre-trained our model on the ImageNet-1k (IN1K) training set. Unlike the previous SSL methods, we additionally input the guidance to provide the region information required for image matting. Referring to most of the previous matting works, we decided to take widely adopted trimap as the guidance, which delineates the background, foreground, and unknown region.

In the pre-training process, the trimap and the corresponding alpha matte are randomly generated first, meanwhile, the two natural images are randomly selected as

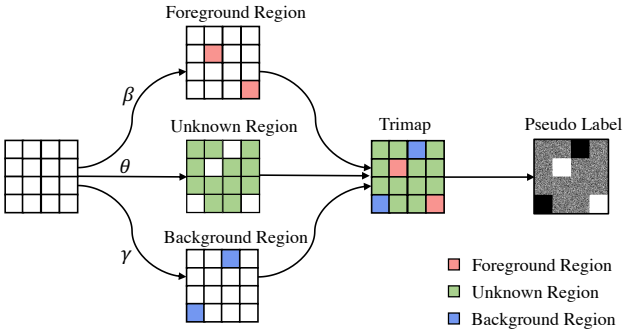

Figure 3. The generation of trimap and pseudo label. θ percent of patches in trimap as unknown region, β percent for background γ percent for foreground, where $\theta+\beta+\gamma = 1$. Then a random alpha matte α is generated as a pseudo label according to the trimap.

background and foreground, respectively. Then the fused image is composited with the alpha matte, background image, and foreground image via Eq.1. At last, the fused image and the one-hot trimap are concatenated as the input of the network to predict alpha matte. The randomly generated alpha matte will be used to supervise the predicted alpha matte. After self-supervised pre-training, we fine-tune on downstream matting dataset using the trained model parameters as initialization.

Next, we will introduce the details of generating the trimap, the corresponding alpha matte, and the objective function.

Trimap generation Trimap, as the crucial guidance for matting, is defined in advance. In this paper, we adopt a very simple but effective method to generate trimap. A blank trimap $T \in \mathbb{R}^{H\times W}$ is created first, where H and W are the height and width, respectively. The trimap is with the same resolution as the input image. Then T will be divided into grids of the same size, such as 7×7 pixels. We randomly select θ percent of grids as the unknown region, β percent of grids as the foreground region, and γ percent of grids as the background region, $\theta + \beta + \gamma = 1$. The background is assigned 0, the foreground is assigned 2, and the unknown region is assigned 1. In the training step, the $\boldsymbol{T}$ is converted into a one-hot embedding as the input guidance of the model. The trimap generation is random and is not independent of the content of the input image. Here, we randomly select based on grids rather than pixels to keep the structure of foreground and background.

Alpha matte generation After obtaining trimap, we continue to generate the corresponding alpha matte α. Alpha matte is a matrix that has the same size as the trimap. Each value α_i of the alpha matte is in the range of 0 to 1 and will be used to fuse the foreground image and the background image. According to Eq. 1, $\alpha = 0$ means the fused image is equal to the background image, $\alpha = 1$ means the fused image is equal to the foreground image. For each position i in the alpha matte, we set $\alpha_i = 1$ if the given trimap $T_i = 2$, and $\alpha_i = 0$ if the given trimap $T_i = 0$. For the position in which $T_i = 1$, we randomly assign a number between 0 and 1 for each pixel. It should be emphasized that, in order to be closer to the annotation of the matting dataset, we randomly select an integer value between 0 to 255 for each pixel in the unknown region, then divide by 255 as the final value. The flow chart is shown in Figure 3. In the experiment section, we discuss different ways to generate trimap and Alpha matte.

Loss function We adopt three kinds of generic loss functions: L1 regression loss, Composition loss [43] and Laplacian loss [17] for both pre-training and fine-tuning stage. The final loss is the sum of the above three losses.

$$L_{final} = L_{l1} + L_{comp} + L_{lap}, \tag{2}$$

where L_{l1} is the absolute difference between the predicted alpha matte and ground truth. The L_{comp} is the absolute difference between the original image and fused image, where the fused image is generated with the original foreground, original background and predicted alpha matte according to Eq. 1. The L_{lap} calculates differences of two Laplacian pyramid representations between ground truth and predicted alpha matte. All the above losses are calculated only in the unknown region.

Discussion Our DPT is similar to natural image matting. We both use additional guidance as an external input to extract the foreground from the fused image. However, there are still some differences. The normal image matting task has a small number of annotations and increases the amount of training data with synthetic images. The trimap is produced from the ground truth alpha matte by binarizing the foreground objects with a threshold with random dilation. In this case, the unknown region is around the boundaries of the foreground object, which also makes it easier for the model to converge. The training of DPT is based on large-scale unlabeled data. Trimap and alpha matte are randomly generated, which increases the difficulty of training. However, the huge amount of unlabeled data makes up for this shortcoming and greatly improves the final performance.

4. Experiments

In this section, we report the detailed setting of our experiments and conduct our experimental evaluations on image matting datasets. In the following, we first compare our method with other state-of-the-art matting methods on Composition-1k [43] and Distinct-646 [30] datasets. Then we compare different alternative pre-training methods with the proposed DPT. We also apply our DPT pre-training on

different backbones to validate its robustness. Extensive ablation studies are conducted to validate the contributions of each technical detail.

4.1. Experimental setting

Data augmentation We do self-supervised pre-training on the ImageNet-1k (IN1K) training set [9]. Following previous pre-training approaches [14, 25], we use the default image input size of 224×224. For each foreground image, we randomly select an image as the background. The trimap and pseudo label are kept the same size as the image. The patch size within the trimap is set as 7×7. The ratios of an unknown region, foreground, and background are set to 75%, 12.5% and 12.5%, respectively. We first perform an affine transformation with a random degree, scale, shear, and flip. After that, we randomly change the Hue values of the image. Finally, we composite the foreground and background images with pseudo alpha matte.

In the fine-tuning stage, we perform our experiments on image matting datasets. The data augmentation is set following MG Matting [45]. Two foreground images are randomly fused with alpha matte, followed by random affine transformations and colour jitterings. Patches with size 512× 512 are cropped around the unknown area in the central area of the foreground image. The augmented foreground and background are then fused according to the ground truth alpha matte.

Pre-training We employ the AdamW optimizer with β_1=0.9 and β_2=0.95 for our objective functions. The learning rate adopts a cosine annealing strategy with the base learning rate of 3×10^{-4} and a weight decay of 0.05. By default, we perform self-supervised pre-training with batch size 512 on a single machine equipped with eight NVIDIA V100 GPUs. The model is trained for 100 epochs with a 10-epoch linear warm-up stage.

Fine-tuning In the fine-tuning stage, we follow the same setting as in MatteFormer [29]. We initialize the network with the Tiny model of Swin Transformer [25] pretrained on IN1K. The input size of the network is 512× 512, batch size of 40 for four V100 GPUs on one machine. We initialize the base learning rate with 10^{-3}. Adam optimizer is adopted with β_1=0.5 and β_2=0.999 for training 200k iterations. In the first 5k iterations, we warm up the learning rate by linear increasing to help the model convergence.

Evaluation metrics Following the common practice in previous matting methods [29, 30, 45], we adopt the sum of absolute differences (SAD), mean squared error (MSE), gradient (Grad), and connectivity errors (Conn) as our evaluation metrics. The fused image and 3-channel trimap are concatenated in the channel dimension as inputs for the networks. The input image is padded to a size of multiple of 32 to facilitate the downsampling of the model and afterwards restored to its original size for evaluation. Note that lower values of the four evaluation metrics indicate higher performance (*i.e*. more accurate alpha matte estimation).

Table 1. Quantitative fine-tuning results on Composition-1K. * indicates additional semantic information as input.

Method	SAD↓	MSE (10^{-3})↓	Grad↓	Conn↓
KNN-Matting [5]	175.4	103.0	124.1	176.4
DIM [43]	50.4	14.0	31.0	50.8
AlphaGAN [27]	52.4	30.0	38.0	-
IndexNet [26]	45.8	13.0	25.9	43.7
SampleNet [37]	40.4	9.9	-	-
Context-Aware [17]	35.8	8.2	17.3	33.2
GCA [30]	35.3	9.1	16.9	32.5
HDMatt [44]	33.5	7.3	14.5	29.9
TIMI-Net [24]	29.1	6.0	11.5	25.4
MG Matting [45]	32.1	7.0	14.0	27.9
SIM* [36]	28.0	5.8	10.8	24.8
RMat [8]	25.0	-	9.0	-
MatteFormer [29]	23.8	4.0	8.7	18.9
TransMatting [3]	26.8	5.2	10.6	22.1
DPT (Ours)	**21.0**	**3.1**	**7.0**	**15.9**

4.2. Quantitative comparison to state-of-the-arts

Composition-1k Here we test our DPT on Composition-1k [43] test set and compare it with state-of-the-art approaches. Composition-1k is a synthetic matting dataset that contains 431 foreground objects and corresponding labelled alpha matte for training. The test set contains 50 foreground objects that are composited with 20 background images chosen from Pascal-VOC [11], a total of 1,000 samples. The proposed DPT surpasses previous methods with a large margin as shown in Table 1. Surprisingly, our DPT outperforms MatteFormer by a large margin, by simply replacing its pre-training weights from IN1K supervised one to our proposed self-supervised one, suggesting the effectiveness of the proposed approach.

Distinct-646 Distinct-646 [30] is a matting benchmark dataset containing 59.6k training images and 1k test images, in total 646 distinct foreground alpha mattes. Unlike the previous matting dataset, it does not provide trimap annotations or other guidance, hence it is difficult to make a fair comparison with other methods. Therefore, we generate trimap by randomly dilating alpha mattes from the ground truth alpha matte as done in MG Matting [45].

After obtaining the trimap guidance, we compare with the state-of-the-art methods and report the performance in Table 2, in which the methods marked with * are trained on Composition-1k, while others are trained on Distinct-646. For a fair comparison, here we compare with MG Matting and MatteFormer by using the exact same testing setting and test on the Distinct-646 test set. It can be seen from the

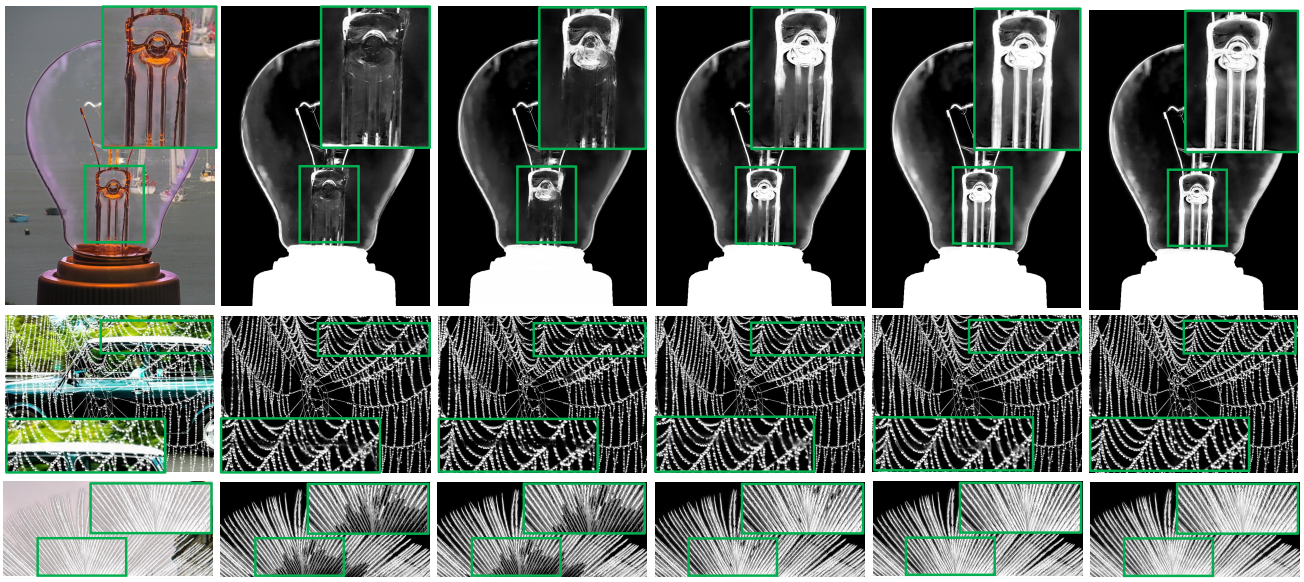

Figure 4. The qualitative results of ours and other state-of-the-art methods on Composition-1k test set.

Table 2. Quantitative results on Distinct-646. Results with * are only trained on Composition-1k, others are trained on Distinct-646 as reported in HAtt Matting [30].

Method	SAD↓	MSE (10^{-3})↓	Grad↓	Conn↓
Closed-Form [19]	105.7	23.0	91.8	114.6
Learning Based [46]	105.0	21.0	94.2	110.4
KNN Matting [5]	116.7	25.0	103.2	121.5
DIM [43]	47.6	9.0	43.3	55.9
HAttMatting [30]	49.0	9.0	41.6	50.0
DIM* [43]	48.7	11.2	42.6	50.0
Index* [26]	47.0	9.4	40.6	46.8
Context-Aware* [17]	36.3	7.1	29.5	35.4
GCA* [30]	39.6	8.2	32.2	38.8
MG Matting* [45]	36.9	6.3	35.0	22.3
MatteFormer* [29]	31.0	4.9	22.6	19.6
DPT(Ours) *	**28.5**	**3.8**	**18.0**	**19.0**

results that the proposed DPT outperforms other methods with a clear margin across all four evaluation metrics.

4.3. Qualitative performance

The qualitative results of ours and other state-of-the-art methods on the Composition-1k test set are shown in Figure 4, in which we compare with GCA [30], MG Matting [45] and MatteFormer [29]. In most samples, various methods can achieve good results, but in some challenging cases (*e.g.* cobweb, light bulb), our method performs much better, especially in detailed regions.

4.4. Analysis on alternative pre-training methods

To validate the effectiveness of the proposed pre-training approach, we consider several representative existing pre-trained models developed for Swin Transformer (Tiny model). The classification supervised pre-training and Transformer-SSL are pre-trained for 300 epochs, SimMIM and DPT are trained for 100 epochs. All the methods are pre-trained on IN1K [9] only, and the pre-trained models are used as initialisation of MatteFormer for fine-tuning on Composition-1k. The results are shown in Table 3:

- *Random*: without any pre-trained weights for initialisation, we randomly initialize the model parameters, and directly train from scratch.
- *Supervised*: the network initializes with the most commonly used IN1K [9] fully-supervised pre-trained model.
- *SimMIM* [42]: a simple framework for SSL by masked image modeling.
- *MoBY* [41]: an SSL method combined with contrastive loss, serving as a representative contrastive learning-based work.

It can be seen from the experimental results that the performance of the model without the pre-trained weight is much worse than those with the pre-trained weight, suggesting that due to the limited number of labelled matting images, prior knowledge needs to be obtained and is indeed helpful. By using pre-trained weights (second row) it

Table 3. With different pre-training methods, quantitative results after fine-tuning on Composition-1k test set.

Pre-train Method	SAD↓	MSE (10^{-3})↓
Random	49.2	14.7
Supervised [9]	23.8	4.0
SimMIM [42]	24.6	4.2
Transformer-SSL (MoBY) [41]	24.0	3.9
DPT (Ours)	**21.0**	**3.1**

Table 4. With different backbone pre-training, fine-tuning quantitative results on Composition-1k test set.

Method	Backbone	Init.	SAD↓	MSE (10^{-3})↓
GCA [30]	Resnet-34	Supervised	35.3	9.1
GCA [30]	Resnet-34	DPT	**33.0**	**8.3**
MG Matting [45]	Resnet-34	Supervised	29.3	6.3
MG Matting [45]	Resnet-34	DPT	**27.1**	**5.5**
MatteFormer [29]	Swin-T	Supervised	23.8	4.0
MatteFormer [29]	Swin-T	DPT	**21.0**	**3.1**

achieves a positive performance gain on the matting task. The remaining two self-supervised methods can achieve comparable performance with fully supervised classification. Surprisingly, our proposed DPT achieves much better performance, even outperforming the supervised counterpart. This verifies that the proposed DPT learns better representations for the matting task.

4.5. Analysis on different backbones

Preceding deep learning-based matting approaches can be roughly divided into two categories in terms of backbone architecture: CNN-based and Transformer-based. Most of the matting works use CNN-based backbones, mainly based on the ResNet [16] architecture. With the rise of Swin Transformer [25], there are also some works based on Transformers. To analyze the scalability of DPT, we conduct experiments by pre-training on both backbones and fine-tuning on the same matting dataset.

For CNN-based approaches, we choose GCA [30] and MG Matting [45]. The backbone used by GCA and MG Matting is ResNet-34 [16]. However, the input guidance of MG Matting is a 1-channel mask, different from trimap-based methods. To better compare with it, we replace its input mask with a 3-channel trimap and retrain the model for testing. The corresponding network layer is also modified accordingly. We use the IN1K classification supervised pre-trained model as the baseline. And compare it with GCA and MG Matting by re-training it using our DPT.

On the other hand, we select a state-of-the-art Transformer-based approach, MatteFormer [29]. As above, we also use the weights of classification supervision as the baseline, and re-train MatteFormer with our DPT. We load the two types of pre-training parameters separately and perform fine-tuning on the MatteFormer. The performance comparison is shown in Table 4.

Note that in the fine-tuning stage, only the pre-trained weights have been changed, but we can see a substantial improvement has been achieved on both backbones when using our DPT. This validates the robustness and transferability of the proposed method.

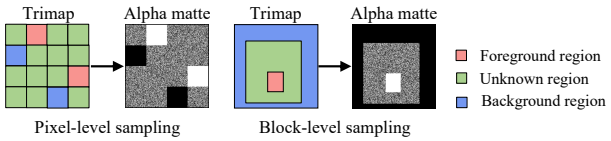

Figure 5. Pseudo alpha matte generation strategy.

Table 5. Results of different pseudo alpha matte strategies.

Strategy	SAD↓	MSE (10^{-3})↓
Pixel-level sampling	21.0	3.1
Block-level sampling	21.0	3.2

Table 6. With different stages of loading DPTpre-trained weights, fine-tuning quantitative results on Composition-1k test set.

Method	Stage	Init.	SAD↓	MSE (10^{-3})↓
MatteFormer	Encoder	Supervised	23.8	4.0
MatteFormer	Encoder	DPT	22.1	3.4
MatteFormer	Encoder+Decoder	DPT	21.0	3.1

4.6. Ablation study

Pseudo alpha matte generation strategy We compare two pseudo alpha matte generation strategies, as illustrated in Figure 5. *1) Pixel-level sampling strategy*. We first divide a 224×224 image into 7×7 patches. For each patch, it is randomly assigned as either foreground, background or unknown regions, and the trimap is generated consequently. Then pseudo alpha matte is randomly generated in each pixel according to the pre-defined trimap.

The results of the two strategies are shown in Table 5. After experiments, both strategies can achieve new state-of-the-art results. We finally chose the pixel-level sampling strategy with slightly higher performance.

Stages of loading pre-trained weights In order to verify the impact of loading pre-trained weights onto different stages on performance, we conduct experiments based on MatteFormer [29]. We compare the supervised method which only loads the pre-trained weights in the encoder, with DPT loading the pre-trained weights in either the encoder or the whole model. As shown in Table 6, our DPT outperforms the baseline by a large margin, which can be further improved when including the pre-trained decoder.

Table 7. The fine-tuning quantitative results on Composition-1k test set, when with and without trimap during pre-training.

Method	Trimap	SAD↓	MSE (10^{-3})↓
MatteFormer	Without	23.4	3.7
MatteFormer	With	21.0	3.1

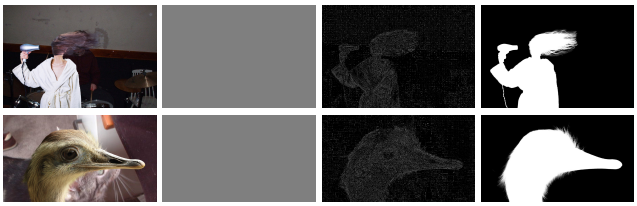

Image Trimap Predicted alpha GT Alpha matte

Figure 6. Qualitative performance of directly predicting alpha matte using our pre-trained model with an *All-unknown* trimap.

Impact of trimap for pre-training Meanwhile, we perform an ablation study to validate the impact of trimap for the pre-training stage. Unlike DPT, the input only contains synthetic images with three channels while the data processing and subsequent fine-tuning stage remain in the same setting. The result is shown as Table 7. Although descent results can be achieved without trimap, adding trimap as an additional input further improves performance significantly.

Contour information validation for pre-training In order to validate the effectiveness of our pre-training approach on learning contour information, we directly applied the pre-trained model on the comp-1k test set, without fine-tuning. We set *all* input trimaps as *unknown*, *i.e*. without any form of guidance. Then we fed the merged image along with such *All-unknown* trimap into the pre-trained model to predict the alpha. As shown in Figure 6, although without any guidance and fine-tuning, our pre-trained model is able to extract object contour information (even hair and fur), suggesting its capacity to learn and understand high-level object information. This study further validates the effectiveness of our approach for the matting tasks.

The ratio of the unknown region in trimap The unknown region is necessary for the loss computation. As the size of the unknown region changes, the final results also change. The foreground and background regions also provide information for the network. Thus a suitable ratio of the unknown region will affect the final performance. Here we analyze different ratios of this unknown region and report the results in Figure 7. It can be seen that when we choose a lower ratio such as 25% or 50%, the results will be slightly worse. If we choose a higher threshold, such as 75% or higher, then the final performance will become stable and get better results. As a result, in this paper, we choose the ratio of 75%.

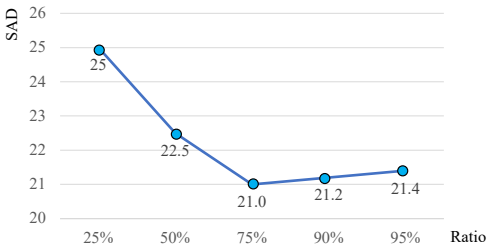

Figure 7. SAD for different ratios of the unknown region in predefined trimap.

Figure 8. Qualitative results of different unknown ratios of trimap. **Top to bottom**: unknown region accounts for 25%, 50% and 75%, respectively. For each image group, from **left to right**: the merged image, trimap, and the foreground image generated by Eq. 1.

Visualization of different ratios of the unknown regions in trimap is shown in Figure 8. As the unknown area increases, the model is supposed to predict more pixels, and the noise points on the prediction result will also increase, which is revealed in the result.

5. Conclusion

In this work, we looked into the fundamental problem of image matting (*i.e*. data) and proposed a disentangled self-supervised pre-training method named DPT. It is designed for the image matting task to enable the utilization of large-scale unlabeled data. More specifically, we generated trimap as an auxiliary input and pseudo alpha matte for supervision, then trained the model towards disentanglement of the fused image. After this pre-training, the derived model was used for the initialization of the downstream image matting task, to boost its performance. Extensive experimental analysis showed the effectiveness and robustness of our DPT. We hope this work could attract attention from the community to think about the data leverage side for the matting task and potentially inspire follow-up research.

Acknowledgements Jianbo Jiao is supported by the Royal Society Short Industry Fellowship SIF\R1\231009.

References

[1] Yağiz Aksoy, Tae-Hyun Oh, Sylvain Paris, Marc Pollefeys, and Wojciech Matusik. Semantic soft segmentation. *ACM Transactions on Graphics (TOG)*, 37(4):1–13, 2018.

[2] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[3] Huanqia Cai, Fanglei Xue, Lele Xu, and Lili Guo. Transmatting: Enhancing transparent objects matting with transformers. *arXiv preprint arXiv:2208.03007*, 2022.

[4] Yue Cao, Zhenda Xie, Bin Liu, Yutong Lin, Zheng Zhang, and Han Hu. Parametric instance classification for unsupervised visual feature learning. *Advances in neural information processing systems*, 33:15614–15624, 2020.

[5] Qifeng Chen, Dingzeyu Li, and Chi-Keung Tang. Knn matting. *IEEE transactions on pattern analysis and machine intelligence*, 35(9):2175–2188, 2013.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[7] Yung-Yu Chuang, Brian Curless, David H Salesin, and Richard Szeliski. A bayesian approach to digital matting. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 2, pages II–II. IEEE, 2001.

[8] Yutong Dai, Brian Price, He Zhang, and Chunhua Shen. Boosting robustness of image matting with context assembling and strong data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11707–11716, 2022.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[12] Marco Forte and François Pitié. f, b, alpha matting. *arXiv preprint arXiv:2003.07711*, 2020.

[13] Eduardo SL Gastal and Manuel M Oliveira. Shared sampling for real-time alpha matting. In *Computer Graphics Forum*, volume 29, pages 575–584. Wiley Online Library, 2010.

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[17] Qiqi Hou and Feng Liu. Context-aware image matting for simultaneous foreground and alpha estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4130–4139, 2019.

[18] Philip Lee and Ying Wu. Nonlocal matting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2193–2200. IEEE, 2011.

[19] Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):228–242, 2007.

[20] Jizhizi Li, Jing Zhang, and Dacheng Tao. Referring image matting. *arXiv preprint arXiv:2206.05149*, 2022.

[21] Yaoyi Li and Hongtao Lu. Natural image matting via guided contextual attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11450–11457, 2020.

[22] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8762–8771, 2021.

[23] Jinlin Liu, Yuan Yao, Wendi Hou, Miaomiao Cui, Xuansong Xie, Changshui Zhang, and Xian-sheng Hua. Boosting semantic human matting with coarse annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8563–8572, 2020.

[24] Yuhao Liu, Jiake Xie, Xiao Shi, Yu Qiao, Yujie Huang, Yong Tang, and Xin Yang. Tripartite information mining and integration for image matting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7555–7564, 2021.

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[26] Hao Lu, Yutong Dai, Chunhua Shen, and Songcen Xu. Indices matter: Learning to index for deep image matting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3266–3275, 2019.

[27] Sebastian Lutz, Konstantinos Amplianitis, and Aljosa Smolic. Alphagan: Generative adversarial networks for natural image matting. *arXiv preprint arXiv:1807.10088*, 2018.

[28] Pedro O O Pinheiro, Amjad Almahairi, Ryan Benmalek, Florian Golemo, and Aaron C Courville. Unsupervised learning of dense visual representations. *Advances in Neural Information Processing Systems*, 33:4489–4500, 2020.

[29] GyuTae Park, SungJoon Son, JaeYoung Yoo, SeHo Kim, and Nojun Kwak. Matteformer: Transformer-based image matting via prior-tokens. In *Proceedings of the IEEE/CVF Con-*

ference on Computer Vision and Pattern Recognition, pages 11696–11706, 2022.

[30] Yu Qiao, Yuhao Liu, Xin Yang, Dongsheng Zhou, Mingliang Xu, Qiang Zhang, and Xiaopeng Wei. Attention-guided hierarchical structure aggregation for image matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13676–13685, 2020.

[31] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[32] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[33] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Background matting: The world is your green screen. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2291–2300, 2020.

[34] Ehsan Shahrian, Deepu Rajan, Brian Price, and Scott Cohen. Improving image matting using comprehensive sampling sets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 636–643, 2013.

[35] Xiaoyong Shen, Xin Tao, Hongyun Gao, Chao Zhou, and Jiaya Jia. Deep automatic portrait matting. In *European conference on computer vision*, pages 92–107. Springer, 2016.

[36] Yanan Sun, Chi-Keung Tang, and Yu-Wing Tai. Semantic image matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11120–11129, 2021.

[37] Jingwei Tang, Yagiz Aksoy, Cengiz Oztireli, Markus Gross, and Tunc Ozan Aydin. Learning-based sampling for natural image matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3055–3063, 2019.

[38] Feng Wang, Huiyu Wang, Chen Wei, Alan Yuille, and Wei Shen. Cp2: Copy-paste contrastive pretraining for semantic segmentation. *arXiv preprint arXiv:2203.11709*, 2022.

[39] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021.

[40] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8392–8401, 2021.

[41] Zhenda Xie, Yutong Lin, Zhuliang Yao, Zheng Zhang, Qi Dai, Yue Cao, and Han Hu. Self-supervised learning with swin transformers. *arXiv preprint arXiv:2105.04553*, 2021.

[42] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022.

[43] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2970–2979, 2017.

[44] Haichao Yu, Ning Xu, Zilong Huang, Yuqian Zhou, and Humphrey Shi. High-resolution deep image matting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3217–3224, 2021.

[45] Qihang Yu, Jianming Zhang, He Zhang, Yilin Wang, Zhe Lin, Ning Xu, Yutong Bai, and Alan Yuille. Mask guided matting via progressive refinement network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1154–1163, 2021.

[46] Yuanjie Zheng and Chandra Kambhamettu. Learning based digital matting. In *2009 IEEE 12th international conference on computer vision*, pages 889–896. IEEE, 2009.