SoftCollage: A Differentiable Probabilistic Tree Generator for Image Collage

Jiahao Yu[1], Li Chen[1]*, Mingrui Zhang[1], Mading Li[2]
[1]School of Software, BNRist, Tsinghua University, Beijing, China
[2]Kuaishou Technology, Beijing, China
{yujh21,zmr20}@mails.tsinghua.edu.cn chenlee@tsinghua.edu.cn limading@kuaishou.com

Abstract

Image collage task aims to create an informative and visual-aesthetic visual summarization for an image collection. While several recent works exploit tree-based algorithm to preserve image content better, all of them resort to hand-crafted adjustment rules to optimize the collage tree structure, leading to the failure of fully exploring the structure space of collage tree. Our key idea is to soften the discrete tree structure space into a continuous probability space. We propose SoftCollage, a novel method that employs a neural-based differentiable probabilistic tree generator to produce the probability distribution of correlation-preserving collage tree conditioned on deep image feature, aspect ratio and canvas size. The differentiable characteristic allows us to formulate the tree-based collage generation as a differentiable process and directly exploit gradient to optimize the collage layout in the level of probability space in an end-to-end manner. To facilitate image collage research, we propose AIC, a large-scale public-available annotated dataset for image collage evaluation. Extensive experiments on the introduced dataset demonstrate the superior performance of the proposed method. Data and codes are available at https://github.com/ChineseYjh/SoftCollage.

1. Introduction

Image collage aims to create a visual summarization with rich information and high aesthetic quality for a group of images. Because this task requires professional collage knowledge, amateurs have a huge demand for automatic image collage tools [16]. Therefore, many research efforts have tried to automate the process of image collage. While many works [4, 12, 18, 19, 25, 26, 33, 41] have achieved a certain level of success in improving visual perception of collage results, they brought about image artifacts [18, 25, 26, 41] and image overlapping [19, 33, 36, 40]. To tackle these defects, some tree-based algorithms [3,8,16,23,37,38] were developed to preserve image content better. A tree-based collage is encoded as a binary tree which leads to a recursive partition of the canvas as illustrated in Fig. 2. In the tree, each leaf node corresponds to an image and each interior node corresponds to a bounding box, whose designation as a horizontal ("H") or vertical ("V") cut corresponds to dividing the box into two child boxes [3]. The existing tree-based methods design a two-stage procedure, where images are arranged in a standard collage tree in the first stage and the tree is mapped to the collage via a specific bijection mapping function in the second stage. Accordingly, the collage layout optimization is cast as an optimal tree structure search problem.

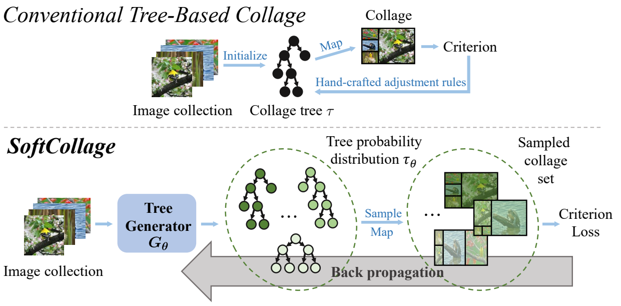


Figure 1. The optimization paradigms of the conventional methods and the proposed *SoftCollage*. We formulates the tree-based collage generation as a differentiable process via softening the discrete tree structure τ into a probability space for the first time. Instead of the hand-crafted adjustment scheme, we directly exploit the gradient of the criterion loss to optimize the tree probability distribution τ_θ, which facilitates the tree structure exploration.

However, all the existing works only resort to heuristic hand-crafted adjustment rules when searching the optimal tree structure, leading to the failure of fully exploring the structure space of collage tree (Fig. 3). Deep learning provides a promising way to learn a high-quality collage tree. Unfortunately, the two-stage tree-based collage generation process is undifferentiable because both stages include dis-

*Corresponding author. This research was partially supported by the National Natural Science Foundation of China (Grant Nos.61972221, 62021002, 61572274) and Tsinghua-Kuaishou Institute of Future Media Data. We thank Xingjia Pan for preparing some comparison results.

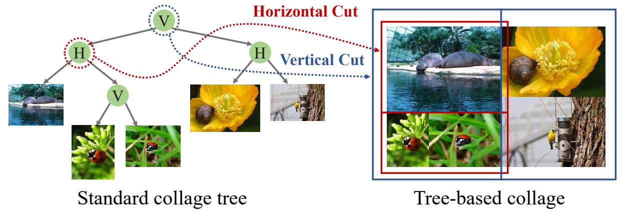


Figure 2. An example of the mapping from a standard collage tree to the tree-based collage.

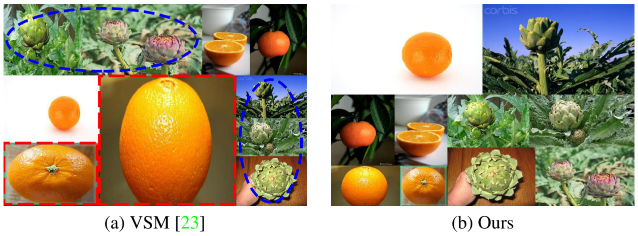
crete operations that prevent back propagation. Although recent tree-based advances [16, 23] utilized learning strategies, they only applied them to yield semantic feature in the first stage so that images with similar features clustered together. These works achieve much improvement because placing correlated images together can facilitate collage informativeness [18, 38, 41]. However, these methods still employed hand-crafted scheme to refine tree structure and failed to fully explore the solution space (Fig. 3). Recently, despite Pan *et al*. [23] introduced back propagation for the first time to fine-tune aspect ratio and splitting ratio, they still failed to propagate the gradients back to optimize the collage tree structure due to the undifferentiable characteristic of the tree-based process.

In this paper, we attack the key problem of differentiating the overall two-stage tree-based collage generation process (Fig. 1). Specifically, firstly we propose a novel neural-based differentiable probabilistic tree generator to model the first stage of tree-based procedure. Our tree generator exploits deep image feature and embedded information including aspect ratio and canvas size to construct a correlation-preserving probabilistic collage tree (PCtree), which builds a probability space via modeling the node type distribution (the cut type of the node is horizontal (“H”) or vertical (“V”)) and the edge connection distribution (the child node is on the left (“L”) or right (“R”)) (Fig. 5). Secondly, we formulate the tree generator optimization as an end-to-end framework resorting to the policy gradient technique [30], which naturally overcomes the differentiation difficulty in the second stage of tree-based procedure. Instead of the hand-crafted adjustment scheme in instance level, our optimization paradigm directly utilizes the gradient of collage criteria loss to optimize the collage tree structure in the level of probability space, which facilitates the exploration of the optimal collage structure.

Furthermore, this field lacks a benchmark dataset with sufficient labels for quantitative evaluation. To facilitate image collage research, we propose AIC, a large-scale public-available annotated dataset for image collage evaluation.

The major contributions can be summarized as follows.

- We propose a novel neural-based probabilistic tree generator which constructs “soft” probabilistic tree structure to build a probability space of correlation-preserving collage tree conditioned on the deep image feature, aspect ratio and canvas size.
- We formulate the tree-based collage generation procedure as a differentiable process for the first time, and introduce an end-to-end learning strategy to perform gradient-based structure optimization.
- We provide a large-scale public-available annotated benchmark dataset for evaluation of image collage method.
- We conduct extensive experiments and user study, and show that our model outperforms the state-of-the-art methods.



(a) VSM [23]

(b) Ours

Figure 3. Due to the failure of fully exploring the structure space of collage tree, the collage generated by the state-of-the-art method (a) still contains images suffering severe aspect ratio distortion (red dotted rectangle) and fails to place similar images together (blue dotted ellipse). Our result (b) preserves aspect ratio and content correlation better.

2. Related Work

Previous works on image collage mainly fall into two categories, *i.e*. parametric method and partitioning-based method. Our tree-based method belongs to the latter.

Parametric methods parameterize a collage with variables including position, scale, orientation and layer index of each image and design well-defined objective functions to solve the optimal variables directly [4, 9, 12, 19, 25–27, 33, 36, 40]. These works either modeled the problem via a probabilistic graphical framework [19, 25, 26, 33, 36, 40] or solved the collage parameters in a heuristic manner [4, 9, 12, 27]. To preserve correlation among images, some methods exploited a feature space to acquire the correlation and projected the images into a visualization space [1, 13, 20, 21, 29, 39]. However, these methods introduce image overlapping and artifact problem.

Partitioning-based methods partition the canvas and assign each image with a corresponding region to compose a collage [3, 8, 10, 16, 18, 23, 28, 31, 37, 38, 41]. Some works utilized Voronoi tessellation [31] and packing algorithm [18, 41] to allocate canvas space for the irregular salient region of each image, which brought about image artifacts when blending image boundaries. Hence, tree-based collage is developed to preserve image content better [3, 8, 16, 23, 28, 37, 38]. Atkins [3] first introduced tree-based collage and solved tree structure in a beam-search

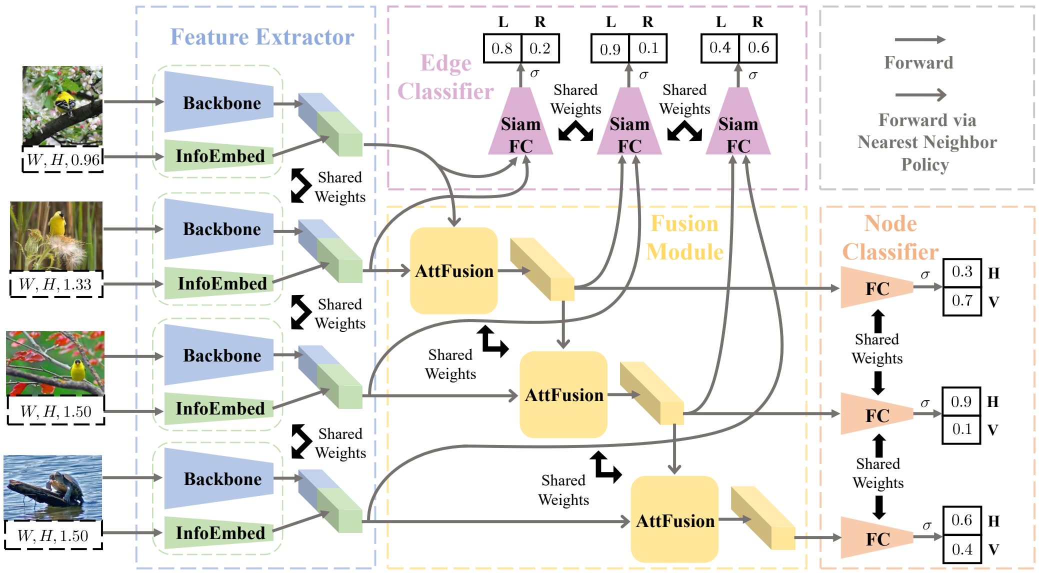


Figure 4. The pipeline of our tree generator. Here the image collection size is four, and our feature extractor initially extracts feature of each image. Subsequently the NNP and fusion module iteratively select child nodes to yield parent feature node in a bottom-up manner until the root feature node of the probability collage tree is acquired. Finally, the edge classifier and node classifier generate $\boldsymbol{p_e}$ and $\boldsymbol{p_n}$ respectively. σ is the softmax activation.

manner. Fan [8] employed genetic algorithm to improve [3] via designing genetic operators of collage tree. Wu and Aizawa [38] initialized tree structure in a greedy manner and adjusted the layout iteratively according to the hand-crafted distortion threshold. These tree-based methods all designed heuristic hand-crafted rules to adjust tree structure, thus failed to fully explore the solution space. Recently, Pan *et al.* [23] utilized back propagation to refine the aspect ratio and splitting ratio of region box in [38]. However, the gradient in [23] still fails to flow back to optimize the tree structure due to the undifferentiable characteristic of the tree-based collage generation process. Different from the prior work, we attack the key problem of differentiating the process via softening the discrete structure of collage tree, and hence our gradient can directly update all the structural details of collage tree.

3. Approach

Problem formulation. According to the literature, a high-quality collage should satisfy the following criteria: 1) Compact. The collage should fully utilize canvas space by blank space minimization. 2) Ratio-preserving. Image in the collage should suffer low aspect ratio distortion to retain the aesthetics. 3) Content-preserving. Image content, especially the salient region, should prevent occlusion. And image overlapping decreases the representativeness and aesthetics of the collage [23]. 4) Correlation-preserving. Recent works show that placing correlated images together facilitates informativeness of the collage [18, 23, 38, 41].

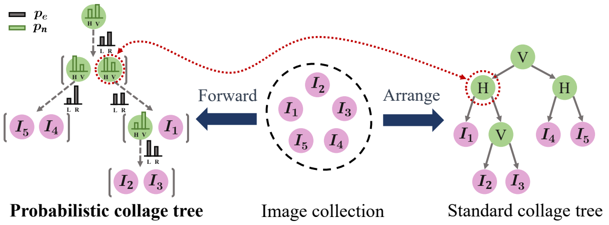


Figure 5. Our probabilistic collage tree softens the standard collage tree structure via modeling the node type distribution as $\boldsymbol{p_n}$ and edge connection distribution as $\boldsymbol{p_e}$.

Therefore, given an image collection $\{I_i\}$,canvas width w and height h, we aim to design a tree generator G. This generator constructs a collage tree τ in the first stage and the tree is mapped to the final collage C via a mapping function g in the second stage. Supposing we integrate the above four criteria into one criterion function F, our goal is to solve the optimal tree generator $G^* = \arg\max_G F\big(g\big(G(w, h, \{I_i\})\big)\big)$.

Overview. To solve the above two-stage problem in an end-to-end manner, firstly we propose a "soft" probabilistic collage tree (PCtree) and design a differentiable tree generator to construct the PCtree. Secondly, we approximate the gradient of criterion loss to optimize our generator via back propagation. These two steps tackle the differentiation problem of the two stages repectively. In the following parts, we firstly present the PCtree, our tree generator and the tree generation algorithm in Sec. 3.1. Afterwards we introduce the model architecture of our neural generator in Sec. 3.2. Finally we present our gradient-based optimiza-

Algorithm 1: Tree construction process

```
Input: $w, h, \{I_i\}$
$N \leftarrow size(\{f_i\})$ ;
$\{f_i\} \leftarrow \{FeatureExtractor(I_i)\}$ ;
repeat
    $f_{n_x}, f_{n_y} \leftarrow NNP(\{f_i\})$ ;
    $f_{n_z} \leftarrow FusionModule(f_{n_x}, f_{n_y})$ ;
    $\boldsymbol{p_e}(n_x, n_y) \leftarrow EdgeClassifier(f_{n_x}, f_{n_y})$ ;
    $\boldsymbol{p_n}(n_z) \leftarrow NodeClassifier(f_{n_z})$ ;
    Remove $f_{n_x}, f_{n_y}$ from $\{f_i\}$ and add $f_{n_z}$ into $\{f_i\}$ ;
    $N \leftarrow N - 1$ ;
until $N = 1$;
```

tion paradigm in Sec. 3.3.

3.1. Probabilistic Collage Tree Generation

Probabilistic collage tree. Standard collage tree represents collage layout using discrete structural parameters including edge connection and node type [3], while the proposed probabilistic collage tree (PCtree) softens the parameters via modeling the node type distribution (the cut type of the node is designated as horizontal ("H") or vertical ("V")) as $\boldsymbol{p_n}$ and the edge connection distribution (the first child node in the child list is designated as the left ("L") or right ("R") child node) as $\boldsymbol{p_e}$, as shown in Fig. 5. The nodes in PCtree and standard collage tree are in one-to-one correspondence. Thus, given an interior node $\tilde{n}$ in a PCtree with child nodes $\tilde{n_i}$ and $\tilde{n_j}$, and the nodes n, n_i and n_j (corresponding to $\tilde{n}$, $\tilde{n_i}$ and $\tilde{n_j}$ respectively) in a standard collage tree, we define $\boldsymbol{p_n}, \boldsymbol{p_e} \in \mathbb{R}^2$ as

$$p_n^{(0)}(\tilde{n}) = p\big(c_n = \text{"H"} | \tau_\theta(\tilde{n_i}), \tau_\theta(\tilde{n_j})\big) \quad (1)$$

$$p_n^{(1)}(\tilde{n}) = p\big(c_n = \text{"V"} | \tau_\theta(\tilde{n_i}), \tau_\theta(\tilde{n_j})\big) \quad (2)$$

$$p_e^{(0)}(\tilde{n_i}, \tilde{n_j}) = p\big(l_n = n_i, r_n = n_j | \tau_\theta(\tilde{n_i}), \tau_\theta(\tilde{n_j})\big) \quad (3)$$

$$p_e^{(1)}(\tilde{n_i}, \tilde{n_j}) = p\big(l_n = n_j, r_n = n_i | \tau_\theta(\tilde{n_i}), \tau_\theta(\tilde{n_j})\big) \quad (4)$$

where $p_n^{(i)}$ and $p_e^{(i)}$ denotes the i-th ($i \in \{0, 1\}$) component of $\boldsymbol{p_n}$ and $\boldsymbol{p_e}$ respectively, c_n is the cut type of n, l_n is the left child node of n, r_n is the right child node of n, and $\tau_\theta(x)$ denotes the subtree of PCtree τ_θ rooted at node x.

Through softening the parameters, we build a probability space for the collage tree and the likelihood of a standard collage tree τ given the PCtree τ_θ can be calculated as

$$p(\tau|\tau_\theta) = \prod_{n \in N(\tau)} p_n^{(\mathbb{1}\{c_n = \text{"V"}\})}(\tilde{n}) \times p_e^{(0)}(\tilde{l_n}, \tilde{r_n}) \quad (5)$$

where $N(\tau)$ is the interior node set of τ, $\tilde{n}$, $\tilde{l_n}$ and $\tilde{r_n}$ denote nodes in the PCtree corresponding to n, l_n and r_n respectively, and $\mathbb{1}\{\cdot\}$ is the indicator function (the value is 1 when the condition is true, otherwise it is 0).

Generator components. To generate the PCtree, we design four learnable components, *i.e.* feature extractor, fusion module, edge classifier and node classifier, as shown in Fig. 4. Feature extractor extracts image semantic features to learn correlation among images and embeds aspect ratio and canvas information to learn layout adjustment. Fusion module fuses the features of child nodes to yield parent node feature for the bottom-up tree construction. Edge classifier determines the edge connection distribution between child nodes and parent node. Node classifier predicts the cut type distribution of interior nodes.

Tree construction algorithm. To preserve correlation among images, we adopt nearest neighbor policy (NNP) to conduct the tree construction in a greedy manner. Given a list of features, our NNP finds the pair of features with the closest Euclidean distance. The tree construction process is described in Algo. 1, where f_n denotes the feature of node n. The time complexity of this algorithm is $O(N^2 \log N)$ with the use of priority queue and hash table, where N is the size of image collection.

3.2. Model Architecture

In this section, we elaborate on the network architecture of our four generator components.

Feature extractor. This component is composed of two-path feature extractors, as shown in Fig. 4. One path employs a pre-trained backbone network to extract content feature $f_{bb}^{(i)}(\theta_{bb})$ from each image I_i and the network parameter θ_{bb} is fine-tuned during training. Another path introduces information embedding $e^{d_w}, e^{d_h}, e^{d_{ar}}$ to inject canvas size and image aspect ratio signals and these signals are fused via a fully connected layer and the ReLU activation function [11] as

$$f_{inf}^{(i)} = ReLU\big(W_1[w \cdot e^{d_w}, h \cdot e^{d_h}, ar_i \cdot e^{d_{ar}}]^T + b_1\big) \quad (6)$$

Here, ar_i is the aspect ratio of image I_i, and we denotes the dimension of $f_{inf}^{(i)}$ and $f_{bb}^{(i)}(\theta_{bb})$ as d_{inf} and d_{bb} respectively. The elements in the embedding row vectors $e^{d_w}, e^{d_h}, e^{d_{ar}}$ are all initialized to one and they are fine-tuned during training. W_1 and b_1 are also learnable parameters. d_w, d_h, d_{ar}, d_{bb} and d_{inf} are hyperparameters.

Because the signals from these two paths are independent, the leaf node feature f_{n_i} of image I_i is obtained via concatenating these two feature vectors.

$$f_{n_i} = concat\big(f_{bb}^{(i)}(\theta_{bb}), f_{inf}^{(i)}\big) \quad (7)$$

Fusion module. This module should obtain the parent feature node via symmetry invariant transforms of the two given child nodes, *i.e.* $f_{fus}(f_{n_i}, f_{n_j}) = f_{fus}(f_{n_j}, f_{n_i})$ where f_{fus} denotes the fusion module. Our idea is to use the self-attentive weighted sum of the two child features to satisfy symmetry invariance. To obtain the weight vectors, we utilize self-attentive embedding technique [17] to design Eq. (10), which injects additive operation into the aspect ratio information fusion process. Moreover, we utilize self-attention mechanism [32] to pre-process the input features for injecting multiplicative signal (Eq. (9)). Bene-

Algorithm 2: Optimization procedure of our model

```
Input: $w, h, \{I_i\}$
Initialize $\theta$ randomly;
$t \leftarrow 0$;
repeat
    Construct probabilistic collage tree $\tau_\theta$ via $\theta$ and $\pi$ in
      accordance with Algo. 1 ;
    Sample $\{\tau_i\}_M$ from $p(\tau|\tau_\theta)$ ;
    Compute $\mathcal{L}(\theta)$ via Eq. (17);
    $\theta \leftarrow \theta - \alpha \times \nabla_\theta \mathcal{L}(\theta)$ ;
    $t \leftarrow t + 1$;
until $t \geq T_m$;
```

fiting from the two-stage transformation, the fusion module is able to memorize a variety of subtree structure schemes, which boosts the learning ability of the model.

$$f_{(i,j)} = [f_{n_i}, f_{n_j}]^T \tag{8}$$

$$f^{'}_{(i,j)} = Attention\big(f_{(i,j)}W_Q, f_{(i,j)}W_K, f_{(i,j)}W_V\big) \tag{9}$$

$$A = softmax\Bigg(W_{s2}\Big(tanh\big(W_{s1}f^{'}_{(i,j)}\big)\Big)\Bigg) \tag{10}$$

$$f_{n_p} = W_{s3}flatten\big(Af^{'}_{(i,j)}\big) + b_2 \tag{11}$$

Here, $W_Q \in \mathbb{R}^{d \times d_Q}, W_K \in \mathbb{R}^{d \times d_K}, W_V \in \mathbb{R}^{d \times d_V}, W_{s1} \in \mathbb{R}^{d_1 \times d_V}, W_{s2} \in \mathbb{R}^{d_2 \times d_1}, W_{s3} \in \mathbb{R}^{d \times d_2 d_V}, b_2$ are all learnable parameters, where d is the dimension of node feature. d_Q, d_K, d_V, d_1 and d_2 are all hyperparameters. Eq. (9) is the scaled dot-product attention parameterized by d_K [32].

Node classifier. A fully connected layer is utilized to model this component as

$$\boldsymbol{p_n}(n) = softmax(W_2 f_n + b_3) \tag{12}$$

where W_2 and b_3 are learnable parameters.

Edge classifier. Different from $\boldsymbol{p_n}$, binary function $\boldsymbol{p_e}$ owns the property that $p_e^{(0)}(n_i, n_j) + p_e^{(0)}(n_j, n_i) = p_e^{(0)}(n_i, n_j) + p_e^{(1)}(n_i, n_j) = 1$, as shown in Fig. 4. Thus, siamese network architecture [5] is employed to model this component as

$$f^{''}_{(i,j)} = W_3 concat(f_{n_i}, f_{n_j}) + b_4 \tag{13}$$

$$f^{''}_{(j,i)} = W_3 concat(f_{n_j}, f_{n_i}) + b_4 \tag{14}$$

$$\boldsymbol{p_e}(n_i, n_j) = softmax\big(f^{''}_{(i,j)}, f^{''}_{(j,i)}\big) \tag{15}$$

where W_3 and b_4 are learnable parameters.

3.3. Gradient-Based Optimization Paradigm

Through building the probability space of collage tree, the tree-based collage generation problem formulation can be modified as solving θ^* subject to $G_{\theta^*} = \arg\max_\theta \mathbb{E}_{\tau \sim L(\tau;\theta,\pi)}\big[F\big(g(\tau)\big)\big]$, where π denotes our NNP, $L(\tau;\theta,\pi) = p\big(\tau|w,h,\{I_i\};\theta,\pi\big) = p(\tau|\tau_\theta)$ and θ is the parameter of tree generator.

Loss function. We define $\mathbb{E}_{\tau \sim L(\tau;\theta,\pi)}\left[F\left(g\left(\tau\right)\right)\right]$ as $\overline{F_\theta}(\tau;\pi)$ and approximate the gradient as

$$\nabla_\theta \overline{F_\theta}(\tau;\pi) \approx \nabla_\theta \left(\frac{1}{M}\sum_{i=1}^{M} F\big(g(\tau_i)\big)\log p(\tau|\tau_\theta)\right) \tag{16}$$

where M is the number of sample τ_i. Therefore, we define the loss function as

$$\mathcal{L}(\theta) = -\frac{1}{M}\sum_{i=1}^{M} F\big(g(\tau_i)\big)\log p(\tau|\tau_\theta) \tag{17}$$

In term of mapping funcition g, We initially utilize an efficient mapping algorithm [8] to generate collage with canvas blank loss r_b, *i.e.* canvas blank space ratio, and we stretch the overall collage to fit the canvas in the post-processing process. Our approach avoids canvas blank space by introducing little aspect ratio distortion. The reason is that canvas blank loss has a significantly worse impact on the user's visual experience than aspect ratio loss, provided that magnitudes of the both losses are similarly small. Moreover, our mapping function benefits from [8] in preventing image content occlusion.

With respect to criterion F, we mainly focus on the ratio preservation criterion because our NNP and mapping function already consider the other three criteria. For this part, we design a reward shaping function R for canvas blank loss r_b as

$$R(r_b) = \begin{cases} -R_0, & r_3 < r_b \\ \frac{R_0(r_b - r_2)}{r_2 - r_3}, & r_2 < r_b \leq r_3 \\ \frac{R_0(\log_{10} r_b - \log_{10} r_2)}{\log_{10} r_1 - \log_{10} r_2}, & r_1 < r_b \leq r_2 \\ R_0, & r_b \leq r_1 \end{cases} \tag{18}$$

where R_0 is the bound of reward value, r_1, r_2 and r_3 are specific blank loss values. The shape design of Eq. (18) is based on the observation that the difficulty of decreasing r_b may be linear in the ratio interval of r_2 to r_3 and it may increase exponentially when r_b is below r_2. R_0, r_1, r_2 and r_3 are hyperparameters. And aesthetics property F_{aes} proposed in [23] is also included in F. Moreover, we design the area penalty F_p to prevent model shrinking some images too much as

$$F_p(C) = -R_0 \times \mathbb{1}\{\exists I \in C \min(h_I, w_I) \leq s_p\} \tag{19}$$

where I is an image in collage C and s_p is a hyperparameter. Therefore, criterion F is defined as

$$F(C) = \lambda_r R\big(r_b(C)\big) + \lambda_a F_{aes}\big(C\big) + \lambda_p F_p\big(C\big) \tag{20}$$

where λ_r, λ_a and λ_p are hyperparameters.

Optimization. Different from hand-crafted adjustment scheme, our optimization paradigm exploits $\nabla_\theta \mathcal{L}(\theta)$ to optimize collage tree probability distribution $p(\tau|\tau_\theta)$ in an end-to-end manner. Algo. 2 shows the optimization paradigm of our model, where T_m is the maximum number of iterations and α is learning rate. At inference stage, optimal collage tree τ^* is determined with maximum likeli-

Theme	Animals	Food	Fruits	Transportation	Sports	Office	Baby	Clothes	Houseware	Instrument	Makeup
Percentage(%)	3.85	11.73	23.22	12.76	4.94	6.44	4.29	18.34	9.38	1.79	3.26

Table 1. The percentage of image number under each theme of ICSS.

hood method as

$$\tau^* = \arg\max_{\tau} p(\tau|\tau_\theta) \tag{21}$$

The detailed derivations in this section is presented in the supplementary materials.

4. Experiments

Baselines. We select three representative tree-based methods as baselines, where one is the state-of-the-art method [23], which is also the mostly related work with ours, and the other two are widely-used commercial softwares [2, 6].

Metrics. We introduce five quantitative metrics to analyze collage results, which are commonly used in state-of-the art works. Among them, three metrics, *i.e.* compactness M_c, ratio preservation M_r and nonoverlapping constraint M_o, are defined identically to [23]. The other two metrics are described as follows.

- Correlation preservation M_n. Gathering correlated images can facilitate informativeness [18, 23, 38, 41].Despite Pan *et al.* [23] considered this end, their metric is actually both an athlete and referee due to the lack of groundtruth label. To tackle this problem, we collected an annotated dataset in Sec. 4.1. Thus, we define $M_n = \frac{1}{N}\sum_I \|P_I - \overline{P}_{c_I}\|_2$, where N is collection size, P_I is the position vector of image I and $\overline{P}_{c_I}$ is the centroid position vector of category label c_I of image I. All position coordinates are normalized by w and h.
- Saliency loss M_s. This metric measures saliency preservation ability. The collage mask is obtained by replacing each image in collage with the corresponding saliency mask. We define $M_s = 1 - |\bigcup_I S_I|/(\sum_I |S_I|)$ where S_I is the saliency mask of image I, $\bigcup_I S_I$ is the collage mask and $|\cdot|$ operator calculates the saliency area of mask.

4.1. Annotated Image Collection Dataset

Collage result for unlabeled image collection cannot support the calculation of M_n and M_s. To encourage research works in this field to compete fairly, we collect an annotated image collection dataset, namely AIC, based on saliency detection dataset DUTS [34] which is partially collected from ImageNet [7] and has high generalization ability [35].

Firstly we select 3402 images from DUTS to build the image collection sampling source, namely ICSS, which covers 72 categories and under each category there is at least 10 images. Subsequently we divide 72 categories into 11 themes manually (Tab. 1). The aspect ratio of images in ICSS ranges from 0.4625 to 1.9048. Each image in ICSS is labeled with category, theme and saliency mask, thus image collection sampled from ICSS is able to support the calculation of M_n and M_s. With the idea of five-fold cross-validation, we divide the images at a ratio of 4:1 in each category into a train set and a test set, and both sets have a near-identical distribution.

Method	Backbone	M_r	M_n	M_s
SHP [6]	-	1.522	0.376	0.239
CLT [2]	-	1.517	0.377	0.232
VSM [23]	-	1.095	0.335	**0**
Ours	ResNet-50 [14]	**1.086**	**0.284**	**0**

Table 2. Quantitative metric results on the train set of AIC.

Method	M_r	M_n
Ours w/o Backbone	1.107	0.379
Ours w/o Info	1.254	0.252
Ours w/o Fusion	30.721	**0.212**
Ours w/o SA	1.503	0.278
Ours (full)	**1.086**	0.284

Table 3. Ablation analysis of our method on the train set of AIC. The first method removes the backbone network and sets $d_{ar} = 1024, d_w = d_h = 32, d_{inf} = 1024$. The second method removes the information embedding in the extractor. The third method replaces the fusion module with feature average operation. The fourth method removes the scaled dot-product attention (Eq. (9)). More results are shown in the supplementary materials.

Method	Pre-trained	Identical theme	Size	M_r	M_n
Ours	✔	✗	$=$	1.155	0.273
Ours	✔	✔	$<$	1.311	0.251
Ours	✔	✔	$>$	1.164	0.254
Ours	✔	✔	$=$	1.091	0.256
Ours	✗	✔	$=$	**1.083**	**0.249**

Table 4. Generalization study of our method on the test set of AIC. The model is directly trained on the test set when not pre-trained. '$>$', '$<$' and '$=$' represent cases where model is pre-trained on a collection of larger, smaller and identical size respectively.

Finally, we develop an image collection sampling framework to generate AIC from ICSS. This framework requires that each image in one collection is sampled from one identical theme of ICSS. Moreover, each collection should include images from at least two categories and each category in collection should have at least two images in order to acquire effective M_n value. Additionally, category distribution of each collection conforms to uniform distribution and is not biased by prior category distribution in ICSS. This framework samples train set and test set of AIC respectively from train set and test set of ICSS. As a result, AIC includes image collections with sizes of 10, 15, 20, 25, 30, 50 and 100. The train set has 562 image collections

5-scale	Excellent (4)	Good (3)	Borderline (2)	Poor (1)	Bad (0)	Score	Kappa
SHP [6]	17.5%	50.8%	27.5%	4.2%	0.0%	2.816	0.82
CLT [2]	16.7%	51.2%	28.8%	3.3%	0.0%	2.813	0.80
VSM [23]	29.2%	**52.9%**	15.4%	2.5%	0.0%	3.088	0.76
Ours	**34.2%**	51.7%	12.0%	2.1%	0.0%	**3.180**	0.80
Side-by-side	Wins	Equally Good	Equally Borderline	Equally Poor	Losses	Δ	Kappa
Ours v.s. SHP [6]	60.6%	28.1%	11.3%	0.0%	0.0%	**60.6%**	0.75
Ours v.s. CLT [2]	63.1%	26.3%	10.6%	0.0%	0.0%	**63.1%**	0.71
Ours v.s. VSM [23]	26.9%	57.5%	9.4%	0.0%	6.2%	**20.7%**	0.67

Table 5. 5-scale human evaluation along with side-by-side human evaluation of collage results on the AIC. The score in 5-scale evaluation is the weighted average. Δ in side-by-side evaluation denotes the gap between the win rate and the lose rate.

Method	Recall	Precision	Accuracy	F1-Score
SHP [6]	0.723	0.625	0.555	0.658
CLT [2]	0.735	0.631	0.564	0.663
VSM [23]	0.808	0.703	0.618	0.745
Ours	**0.865**	**0.771**	**0.669**	**0.810**

Table 6. The results of information conveying test. We investigate four indicators, *i.e*. recall, precision, accuracy and F1-score to evaluate the information conveying ability of collages.

including 18535 images and the test set has 62 image collections including 1260 images. The framework is detailed in the supplementary materials.

4.2. Experiment Settings

Experimental data. We use the train set[*] of AIC for the baseline comparison experiment and the ablation analysis, and the test set for generalization study. The user studies are conducted with the collage results on the train set of AIC.

Implementation details. We implement the proposed framework using the PyTorch toolbox [24] on one GeForce RTX 3090 GPU. We adopt the ResNet-50 [14] pre-trained on the ImageNet [7] as the backbone network in our feature extractor and use the Adam optimizer [15] to train our model for each image collection. The other implementation details are presented in the supplementary materials.

4.3. Quantitative Experiments

Comparison to baseline methods. Comparing to baseline methods, our method achieves similar or better M_r, M_n and M_s metric results on the AIC, shown in Tab. 2. As for M_c and M_o, baseline methods and our model all achieve the optimal zero value due to the advantage of tree-based structure, and thus they are not included in Tab. 2 for conciseness. Fig. 7 shows some comparison results. More results are presented in the supplementary materials.

Ablation analysis. To show the detailed contributions of the components in our model, we conduct ablation experiments on the AIC (Tab. 3). Only M_r and M_n metrics are demonstrated because the other metrics do not change in the ablation. It is shown that the backbone network and the information embedding in the feature extractor are effective in preserving correlation and reducing ratio distortion respectively. The results also demonstrate that the fusion mudule is critical to the learning ability of our model, without which our model can only yield vertical cut in collage (see Fig. 6) and thus produces bad results. Moreover, the self-attention mechanism improves our model much due to the injection of multiplicative operation of aspect ratio information.

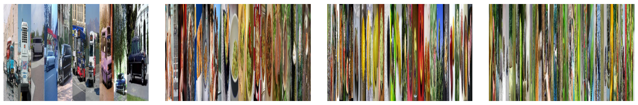


Figure 6. Replacing the fusion module of our generator with feature average operation results in collages with only vertical cut.

Generalization study. Different from the prior work, our generator can learn layout knowledge during optimization and generalize to other collection without training. To study the bottleneck data factors that impact the model generalization ability, we conduct an analysis via controling variates of theme and collection size, shown in Tab. 4. The pre-trained collections are randomly selected as long as they satisfies the corresponding conditions. Tab. 4 shows that the size of pre-trained collection has more significant impact on model generalization ability than the theme of that.

4.4. User Studies

Besides the quantitative measures, we conducted two user studies to evaluate the effectiveness of our method. We select 16 image collections for this stage, which cover all sizes and themes of the collections in the AIC. Each user study was conducted with different groups of participants via different questionnaire, and collages in each questionnaire were ordered in a random way to avoid biasing judges.

Human evaluation. Firstly we carried out the 5-scale evaluation. To measure the gain in our method over the baselines, we also conducted the side-by-side evaluation. This comparative task is easier than 5-scale rating task for human and thus can produce more reliable results. Additionally, Fleiss' Kappa score is used to gauge the reliability of the agreement between evaluators. The details of these evaluations are presented in the supplementary materials. The

[*] We learn a specific generator in each image collection respectively in the train set. Thus, our train set is different from the definition of train set in the traditional deep learning context.

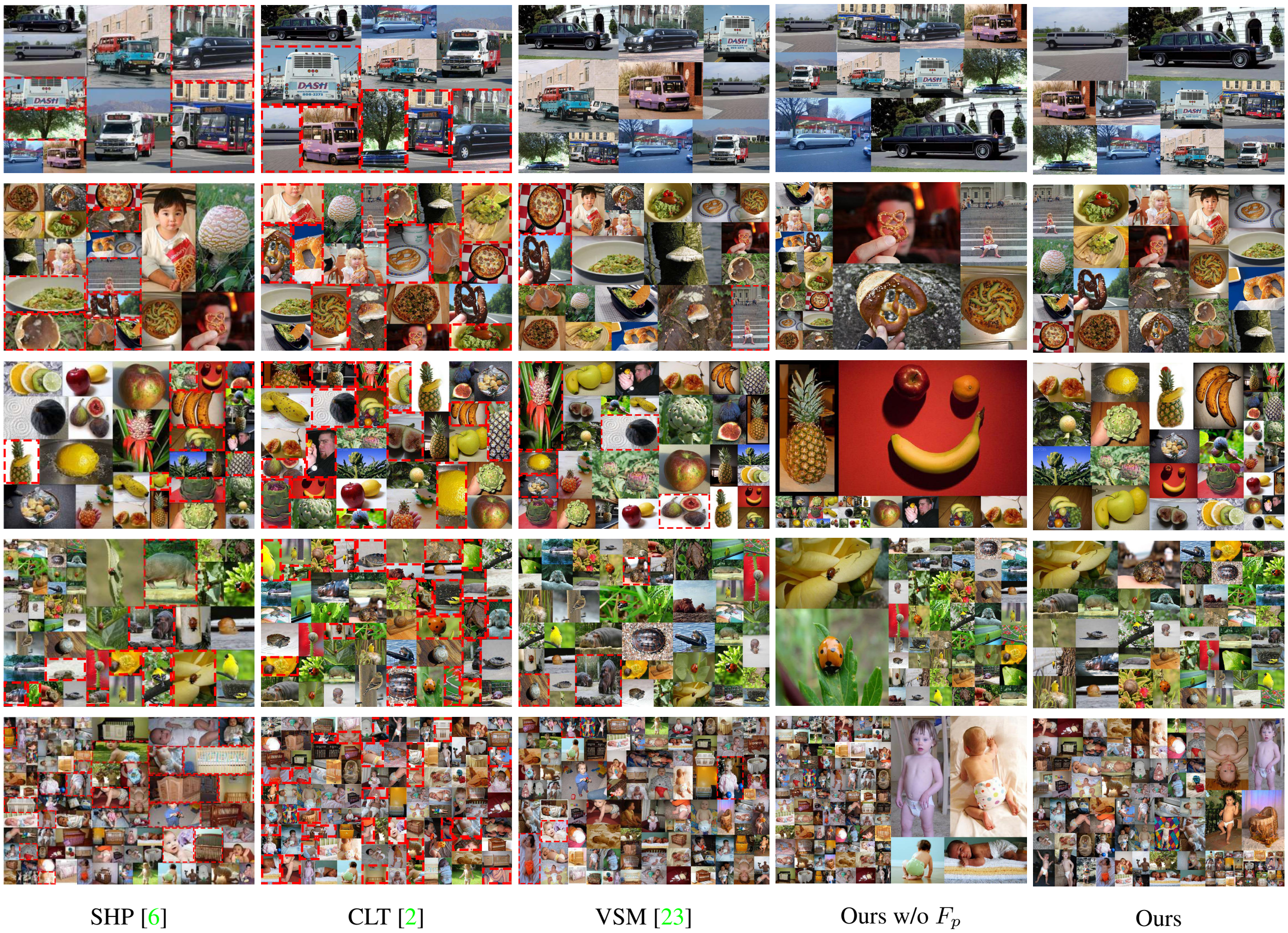


Figure 7. Comparison of the collage results generated by different methods on the AIC. We can see that SHP [6] and CLT [2] both introduce content occlusion (red dotted rectangle) into the images in collage. Despite VSM [23] circumvents this defect, the results still contain images suffering high aspect ratio distortion (red dotted rectangle), particularly when the image collection size is large. However, our method takes advantage of the probability space to produce results closer to the global optimal. Notably employing loss function without F_p of Eq. (19) to train our model leads to drastic imbalance in image area assignment in collages.

results, illustrated in Tab. 5, suggest that our method is substantially superior to all baselines in producing high-quality collage from human's perspective. The high Kappa scores imply that a major agreement prevails among the evaluators.

Information conveying test. We further validate the effectiveness of our NNP via the information conveying test according to [22, 23]. Twenty subjects participated in the test and they were equally divided into four groups. Each group corresponds to one collage method. For each image collection, we showed participants the corresponding collage for 20 s and then asked them to perform a binary classification test, namely selecting the images that they had seen in the collage, on an image set including five groundtruth images and five negative samples (sharing the identical theme with the groundtruths). Tab. 6 shows the test results. Our collage benefits from the NNP and thus outperforms the other baselines. We find that the images selected by participants account for approximately 72%, which implies that participants are inclined to choose more images as remembered, leading to a higher recall than precision.

5. Conclusion

In this paper, we present *SoftCollage*, a novel tree-based collage method. Our key idea is to soften the discrete tree structure into the probability space. By modeling the conditional probability distribution of collage tree via the proposed tree generator, we can formulate the collage generation as a differentiable process and optimize the layout with the gradient of criterion loss instead of the hand-crafted adjustment scheme. We demonstrate the effectiveness of our method via extensive experiments on the proposed large-scale dataset AIC. Currently, the GPU memory consumption of our model is high when the size of image collection is large. Because of the extensibility of our method in model architecture design, in the future we will explore the lightweight design and knowledge distillation of our model.

References

[1] Similarity preserving snippet-based visualization of web search results. *IEEE transactions on visualization and computer graphics*, 20(3):457–470, 2014. 2

[2] Collageit. online, 2019. https://www.collageitfree.com/. 6, 7, 8

[3] C Brian Atkins. Blocked recursive image composition. In *Proceedings of the 16th ACM international conference on Multimedia*, pages 821–824, 2008. 1, 2, 3, 4

[4] Simone Bianco and Gianluigi Ciocca. User preferences modeling and learning for pleasing photo collage generation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 12(1):1–23, 2015. 1, 2

[5] Jane Bromley, James W Bentz, Léon Bottou, Isabelle Guyon, Yann LeCun, Cliff Moore, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(04):669–688, 1993. 5

[6] V. Cheung. Shape collage. online, 2013. http://www.shapecollage.com/. 6, 7, 8

[7] 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. 6, 7

[8] Jian Fan. Photo layout with a fast evaluation method and genetic algorithm. In *2012 IEEE International Conference on Multimedia and Expo Workshops*, pages 308–313. IEEE, 2012. 1, 2, 3, 5

[9] Yuan Gan, Yan Zhang, Zhengxing Sun, and Hao Zhang. Qualitative photo collage by quartet analysis and active learning. *Computers & Graphics*, 88:35–44, 2020. 2

[10] J. Geigel, A. Loui, and E. Loui. Automatic page layout using genetic algorithms for electronic albuming. *Proceedings of SPIE - The International Society for Optical Engineering*, pages 79–90, 2001. 2

[11] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. *Journal of Machine Learning Research*, 15:315–323, 2011. 4

[12] Stas Goferman, Ayellet Tal, and Lihi Zelnik-Manor. Puzzle-like collage. In *Computer graphics forum*, volume 29, pages 459–468. Wiley Online Library, 2010. 1, 2

[13] E. Gomez-Nieto, W. Casaca, D. Motta, I. Hartmann, G. Taubin, and L. G. Nonato. Dealing with multiple requirements in geometric arrangements. *IEEE Transactions on Visualization & Computer Graphics*, 22(3):1223–1235, 2016. 2

[14] 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. 6, 7

[15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 7

[16] Yuan Liang, Xiting Wang, Song-Hai Zhang, Shi-Min Hu, and Shixia Liu. Photorecomposer: Interactive photo recomposition by cropping. *IEEE transactions on visualization and computer graphics*, 24(10):2728–2742, 2017. 1, 2

[17] Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 4

[18] Lingjie Liu, Hongjie Zhang, Guangmei Jing, Yanwen Guo, Zhonggui Chen, and Wenping Wang. Correlation-preserving photo collage. *IEEE transactions on visualization and computer graphics*, 24(6):1956–1968, 2017. 1, 2, 3, 6

[19] Tie Liu, Jingdong Wang, Jian Sun, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Picture collage. *IEEE Transactions on Multimedia*, 11(7):1225–1239, 2009. 1, 2

[20] G. P. Nguyen and M. Worring. Interactive access to large image collections using similarity-based visualization. *Journal of Visual Languages & Computing*, 19(2):203–224, 2008. 2

[21] E. G. Nieto, W. Casaca, L. G. Nonato, and G. Taubin. Mixed integer optimization for layout arrangement. In *Graphics, Patterns & Images*, 2013. 2

[22] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001. 8

[23] Xingjia Pan, Fan Tang, Weiming Dong, Chongyang Ma, Yiping Meng, Feiyue Huang, Tong-Yee Lee, and Changsheng Xu. Content-based visual summarization for image collections. *IEEE transactions on visualization and computer graphics*, 2019. 1, 2, 3, 5, 6, 7, 8

[24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019. 7

[25] Carsten Rother, Lucas Bordeaux, Youssef Hamadi, and Andrew Blake. Autocollage. *ACM transactions on graphics (TOG)*, 25(3):847–852, 2006. 1, 2

[26] Carsten Rother, Sanjiv Kumar, Vladimir Kolmogorov, and Andrew Blake. Digital tapestry [automatic image synthesis]. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 589–596. IEEE, 2005. 1, 2

[27] M. Shuang and W. C. Chang. Automatic creation of magazine-page-like social media visual summary for mobile browsing. In *2016 IEEE International Conference on Image Processing (ICIP)*, 2016. 2

[28] Yu Song, Fan Tang, Weiming Dong, Feiyue Huang, Tong-Yee Lee, and Changsheng Xu. Balance-aware grid collage for small image collections. *IEEE Transactions on Visualization and Computer Graphics*, 2021. 2

[29] Hendrik Strobelt, Marc Spicker, Andreas Stoffel, Daniel Keim, and Oliver Deussen. Rolled-out wordles: A heuristic method for overlap removal of 2d data representatives. In

Computer Graphics Forum, volume 31, pages 1135–1144. Wiley Online Library, 2012. 2

[30] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000. 2

[31] Li Tan, Yangqiu Song, Shixia Liu, and Lexing Xie. Imagehive: Interactive content-aware image summarization. *IEEE computer graphics and applications*, 32(1):46–55, 2011. 2

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 4, 5

[33] Jingdong Wang, Long Quan, Jian Sun, Xiaoou Tang, and Heung-Yeung Shum. Picture collage. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 347–354. IEEE, 2006. 1, 2

[34] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 136–145, 2017. 6

[35] Wenguan Wang, Qiuxia Lai, Huazhu Fu, Jianbing Shen, Haibin Ling, and Ruigang Yang. Salient object detection in the deep learning era: An in-depth survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 6

[36] Yichen Wei, Yasuyuki Matsushita, and Yingzhen Yang. Efficient optimization of photo collage. *Microsoft Research, Redmond, WA, USA, MSRTR-2009-59*, 2009. 1, 2

[37] Zhipeng Wu and Kiyoharu Aizawa. Picwall: Photo collage on-the-fly. In *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–10. IEEE, 2013. 1, 2

[38] Zhipeng Wu and Kiyoharu Aizawa. Very fast generation of content-preserved photo collage under canvas size constraint. *Multimedia Tools and Applications*, 75(4):1813–1841, 2016. 1, 2, 3, 6

[39] Xintong, Han, Chongyang, Zhang, Weiyao, Lin, Mingliang, Xu, Bin, and Sheng. Tree-based visualization and optimization for image collection. *IEEE transactions on cybernetics*, 46(6):1286–300, 2016. 2

[40] Yingzhen Yang, Yichen Wei, Chunxiao Liu, Qunsheng Peng, and Yasuyuki Matsushita. An improved belief propagation method for dynamic collage. *The Visual Computer*, 25(5):431–439, 2009. 1, 2

[41] Zongqiao Yu, Lin Lu, Yanwen Guo, Rongfei Fan, Mingming Liu, and Wenping Wang. Content-aware photo collage using circle packing. *IEEE transactions on visualization and computer graphics*, 20(2):182–195, 2013. 1, 2, 3, 6