Aesthetic Photo Collage with Deep Reinforcement Learning

Mingrui Zhang, Mading Li, Jiahao Yu, Li Chen

***Abstract*—Photo collage aims to automatically arrange multiple photos on a given canvas with high aesthetic quality. Existing methods are based mainly on handcrafted feature optimization, which cannot adequately capture high-level human aesthetic senses. Deep learning provides a promising way, but owing to the complexity of collage and lack of training data, a solution has yet to be found. In this paper, we propose a novel pipeline for automatic generation of aspect ratio specified collage and the reinforcement learning technique is introduced in non-content-preserving collage. Inspired by manual collages, we model the collage generation as a sequential decision process to adjust spatial positions, orientation angles, placement order and the global layout. To instruct the agent to improve both the overall layout and local details, the reward function is specially designed for collage, considering subjective and objective factors. To overcome the lack of training data, we pretrain our deep aesthetic network on a large scale image aesthetic dataset (CPC) for general aesthetic feature extraction and propose an attention fusion module for structural collage feature representation. We test our model against competing methods on movie and image datasets and our results outperform others in several quality evaluations. Further user studies are also conducted to demonstrate the effectiveness.**

***Index Terms*—aesthetic assessment, photo collage, reinforcement learning**

I. Introduction

With the rapid development of the Internet, there has been an increasing popularity of multimedia. The drastic increase in images and videos has led to information explosion, and how to efficiently display diverse content to users within limited spaces has become a popular topic.

Photo collage has been proposed to automatically arrange multiple images on a given canvas. It is widely used for various purposes, such as advertising and photo summarization. Yet the scalability and flexibility also make it a challenging task to generate photo collage with high aesthetic quality.

Many studies have addressed this issue in the past decades, and they can be broadly categorized into content-preserving and non-content-preserving collages. Content-preserving collages[1], [2], [3], [4], [5] are dedicated to arranging photos in accordance with content relevance without cropping or occlusion. The typical procedure includes the selection of representative photos and automatic layout generation. The primary optimization goal is the collage layout. Non-content-preserving collages[6], [7], [8], [9], [10], [11], [12], [13], [14] aim to maximize the information conveyed in the final collage. Informativeness of the results is achieved by overlaying or cropping the irrelevant area. Several derived styles of non-content-preserving collages, including region-partitioned [6], [8], [9], overlay [11], [12], [14], and blending [10], [13] styles, have emerged due to the flexible definition.

To achieve automatic photo collage, previous works mainly exploit handcrafted-based features. Customized features typically include salience, color, and texture and are developed based on a single image. However, such measures cannot always provide adequate collage representation, which is a fundamental feature of collages, because they ignore the composition and collocation details between adjacent images. Deep learning has been used for comprehensive feature representation to overcome the limitations of handcrafted features. However, due to the subjectivity and complexity of collage tasks, training data are lacking and unsuitable for supervised learning. Moreover, given that photo collage generation is a multi-step task, it increases the infeasibility of directly applying deep learning, and traditional methods based on one-stage optimization of state variables cannot always generate high-quality collages.

Motivated by these challenges, we propose a novel pipeline for automatic photo collage generation. Inspired by manual collages, we decompose the collage generation task into interpretable steps and model it as a reinforcement learning (RL) process.As illustrated in Fig. 1, the proposed model includes the deep aesthetic network and the collage generation module.

The deep aesthetic network is composed of the main aesthetic network and the attention fusion module. Compared with traditional non-content-preserving collages that consider ”single images” as basic components and maximize the salience information, the deep aesthetic network formulates ”adjacent images” as basic components and aims to maximize the aesthetic quality from sub-regions. The main aesthetic network extracts general aesthetic features and the attention fusion module is designed to extract the complex structural features of a photo collage, such as the composition and collocation of different sub-images. Specifically, for effective adoption in photo collages, the attention fusion module uses multi-patch information with an attention mechanism to effectively represent the complex features of a collage [15], [16], [17].

Given that manual annotations for photo collages require highly skilled designers and because photo collages are complex and need a substantial amount of training data, the high

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. *(corresponding author: Li Chen)*

Mingrui Zhang, Jiahao Yu and Li Chen are with the School of Software, BNRist, Tsinghua University, Beijing, China. (e-mail: zmr20@mails.tsinghua.edu.cn, yujh21@mails.tsinghua.edu.cn, chenlee@tsinghua.edu.cn)

Mading Li is with the Video Technology Team, Kuaishou, Beijing, China. (e-mail: limading@kuaihsou.com)

TABLE I: Previous studies on photo collages can be broadly categorized into content-preserving collages and non-content-preserving collages.

Category	Subcategory	Basic Operation
Content-Preserving	Tree-based, Power-diagram based	Scaling, Translation
Non-Content-Preserving	Region-partitioning style, Overlay style, Blending style	Scaling, Translation, Cropping, Rotation, Layering

cost involved in constructing a training dataset for photo collages is unrealistic. Hence, we pretrain our aesthetic network on a large-scale image aesthetic dataset (i.e., Comparative Photo Composition or CPC) [18] for general aesthetic feature extraction and propose an attention fusion module for collage feature extraction.

With the aesthetic and structural feature representation from the deep aesthetic network, we formulate collage generation as a sequential decision process and present an improved RL framework. Inspired by the data-driven experience from human collages[12], we design a policy network to manipulate the global layout and local detail properties of individual images, including the orientation angle, spatial position, and placement order. In each step, the policy network makes improvements and generates an aspect-ratio-specified collage, and the value network helps realize stable policy making. The reward design considers subjective and objective factors to instruct the agent to generate collage results with a balanced composition and minimal blank spaces. The policy making is trained with the advantage actor–critic (A2C) algorithm.

Our main contributions are as follows:

- We propose a novel pipeline for automatic photo collage generation. We decompose the collage generation task into interpretable steps and model it as an RL process. Global layout and local detail policies are designed to guide the generation of collages.
- We develop the deep aesthetic network for general aesthetic feature extraction and propose an attention fusion module for structural collage feature representation. The deep aesthetic network formulates "adjacent images" as basic components and maximizes the composition and aesthetic quality from sub-regions. To overcome the lack of collage datasets, we pretrain the main aesthetic network on a large-scale image aesthetic dataset (CPC).
- We evaluate our model against several competing methods on video and image datasets and outperform the others in several quality evaluations. To demonstrate the effectiveness, two user studies are conducted and the correlation between the proposed aesthetic score and the user's opinions is explored.

II. Related Works

Photo collage. Photo collage aims to create a visually appealing summary by arranging multiple images on a given canvas. As summarized in Table I, previous work on photo collages can be broadly categorized into content-preserving collages and non-content-preserving collages.

(a.) Content-preserving collages [1], [2], [3], [4], [5] arrange photos in accordance with the content relevance without cropping or occlusion. Apporaches in [1] and [2] exploited tree-based page division to recursively split a canvas and generated collages in real time. The approach in [3] used a Voronoi tree map and adjusted the layout for image collection. Content-preserving methods are also referred to as photo layout methods and usually specialize in processing large-scale inputs. However, the basic operations are limited and only involve scaling and translation.

(b.) Non-content-preserving collages mainly include region-partitioning-based methods [6], [7], [8], [9] and customized energy terms optimization-based methods[10], [11], [12], [13], [14]. The general goal of region-partitioning-based methods is to allocate space for the irregular salient area on a canvas. The methods of partitioning the canvas into separate disjointed areas mainly include Voronoi tessellation [9], circle packing [7] and feature embedding [8]. However, advance division causes problems when inputted with differently shaped objects that cannot be compactly bounded. Customized energy term optimization-based methods view photo collage generation as an optimization problem with well-defined objective functions. The approach in [10] defined four energy terms to achieve the selection of representative images, seamless blending, and spatially efficient layout, but artifacts exist along the borders. Another idea is the overlay style, which is common in real life and can often be found in albums designed by artists. The approach in [11] defined different energy terms to achieve salience maximization, blank space minimization, and natural preference. The approach in [19] generated collages that focus on the optimization of the overlapping area, layer uniqueness, and angular diversity.

Compared with traditional methods that focus on optimizing regions of interest and salience information, our method adopts the deep learning technique to provide comprehensive collage representation, which helps produce high-quality collages.

Although a recent study [5] has already explored RL generation of content-preserving collages with a handcrafted reward to improve balance, it has four major differences from our work. (1) Unlike feature extraction on "single images", our deep aesthetic network formulates "adjacent images" as basic components and aims to maximize the aesthetic quality from sub-regions. (2) We consider the global layout and local details for the generation of non-content-preserving collages. (3) Instead of a handcrafted reward, our network is trained with an aesthetic reward, which considers subjective and objective factors. (4) We propose the AutoCrop module to inherently generate aspect-ratio-specified collages.

Aesthetic evaluation. Aesthetic evaluation for images has been extensively examined and successfully employed in multiple tasks, such as image quality assessment [20], [21], image

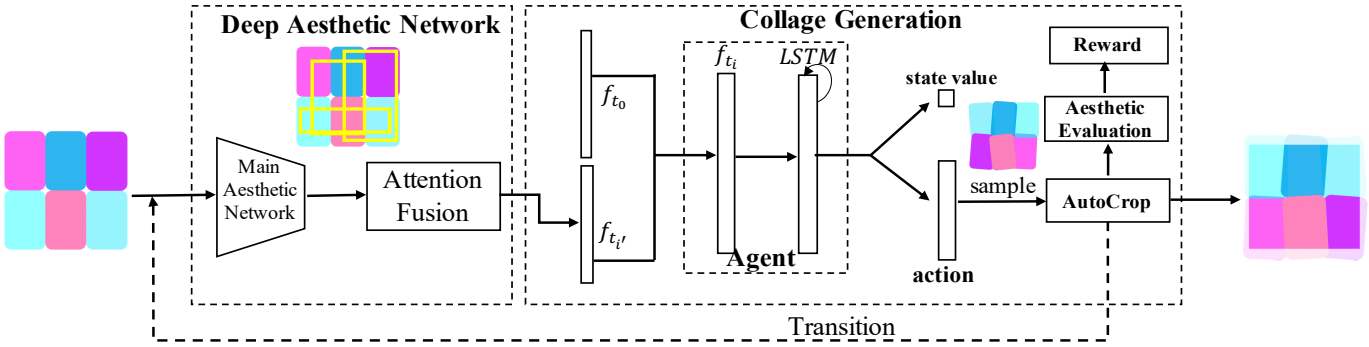


Fig. 1: The proposed network architecture for automatic collage generation. First, the aesthetic network takes in concatenated pictures best suited for the aspect ratio specified canvas as initialization. At each step, the deep aesthetic network first extracts both general aesthetic feature and structural feature representations using the pretrained image aesthetic network and Attention Fusion module for the current collage. Then the current feature concatenated with initial feature is fed into the actor-critic network with action and state output. The agent makes policy based on past observations and samples from the action space to manipulate the global layout of a collage and the local detail properties of individual images as described in Table II. Lastly, the AutoCrop module adapts the generated photo collage with irregular shapes to an aspect ratio-specified canvas, after which the evaluation network computes the aesthetic score and the reward is estimated for current policy.

cropping [22], [23] and image composition [18], [24], and it benefits from the powerful feature representation of deep neural networks. Typical image aesthetic assessment approaches rely on multi-patch representation [15], [16], [17], [25], which represents each image with multiple cropped patches to learn global and local detail information simultaneously, and have been proven useful. Although many succeeding works [17], [25] focus on improvement and further generalization, they remain limited to single images. By contrast, our deep aesthetic network is designed to extract general image aesthetic features and meaningful structural features for a photo collage with the help of the proposed attention fusion module, thus providing a comprehensive photo collage representation.

User preference modeling. To generate collages that satisfy the preference of most people, previous studies have mainly exploited handcrafted standards based on information richness [10], [11], [14], canvas area coverage [11], [14], and information ratio balance [5]. The approach in [12] quantified the criteria and presented seven other subjective preferences through detailed user studies. The new criteria consider global and local details and provide sufficient insights into generating aesthetic collages. On the basis of the experience in user preference modeling, we design our deep aesthetic network to generate collages from the global layout and local details.

RL. RL methods have been applied to multiple computer vision tasks, including image cropping [26], [27], image enhancement [28] image restoration [29] and object tracking [30]. These methods simulate iterative manual modification heuristically, making the operating steps interpretable and easy to understand. Compared with supervised methods, RL-based models do not require heavy annotations and are suitable for subjective tasks, such as collage generation.

III. Methods

Inspired by manual collages, we decompose collage generation into interpretable steps and model it as an RL process. Figure 1 shows that the deep RL model includes a deep aesthetic network for comprehensive feature representation and a collage generation module for agent training.

A. Deep Aesthetic Network for Photo Collage

The deep aesthetic network is designed to extract representative features for a collage. It is composed of the main aesthetic network and the attention fusion module for comprehensive collage feature representation.

Main aesthetic network for general aesthetic feature extraction. Given that a collage is composed of multiple images, encoding information directly from the holistic collage may cause vast information loss and cannot capture local details. Alternatively, the main aesthetic network represents the collage with a bag of predefined patches that densely slides over different scales and aspect ratios of a normalized collage to explore general aesthetic attributes among images.

The network architecture is composed of nine layer convolution blocks and resembles the object detection architecture Single Shot MultiBox Detector [31]. Considering the high expense of manual annotations and the lack of high-quality photo collage datasets, we pretrain our main aesthetic network on a large-scale image aesthetic dataset (CPC) for general aesthetic feature representation.

Attention fusion Module for structural aesthetic feature representation. The components include the fusion module and the attention layer. The fusion module provides aggregated information from orderless patches for effective adaptation in a collage. On this basis, the attention layer is designed for structural aesthetic feature representation of a collage.

Given that the part of a collage of the most concern is the composition and color collocation of composed sub-images, the fusion module shifts its focus to the composition among adjacent sub-images instead of local parts from a single image. Specifically, the fusion module sets standards for feature selection extracted from a candidate patch with an area proportion greater than η to achieved improved composition quality and harmonious content placement. The calculation is as follows:

$$f'_{p_i} = f_{p_i} \cdot \delta\left(\frac{s\left(P_i\right)}{s(C)} > \eta\right) \tag{1}$$

where P_i stands for the i-th patch, $s(\cdot)$ and f represent the area and feature for the i-th patch, respectively; and $\delta(x) = 1$ if x is true. The collage is represented as $f(C) = \left[f'_{p_1}, f'_{p_2}, \ldots, f'_{p_n}\right]$.

The attention layer assigns dynamic weights to features from selected patches for effective learning of the complex structural features of a collage. Similar to handcrafted collages that highlight the important images, we introduce central rule frequently used in photography and pay specific attention to patches close to the center of the collage for further aesthetic quality improvement. Considering that the extracted patch information share the same weights, we inherently add the central rules in the multi-patch fusion process via an attention mechanism, which is formulated as

$$l_i = \|\frac{y_i}{h_c} - y_c\|_2 + \|\frac{x_i}{w_c} - x_c\|_2 \tag{2}$$

$$\alpha_i = \frac{s(P_i)}{s(C)} \cdot (1 - l_i) \tag{3}$$

where (y_i, x_i) denotes the center coordinate of the i-th patch, (h_c, w_c) corresponds to the canvas size, and (y_c, x_c) is set as (0.5, 0,5) to represent the focus center on the canvas. The weighting factors are presented as $\alpha(C) = [\alpha_1, \alpha_2, \ldots, \alpha_n]$. Then, by combining the general and structural aesthetic features, the representation of the holistic collage is computed as:

$$F(C) = \alpha(C) \cdot f(C) \tag{4}$$

B. Collage Generation

We cast the collage generation as a sequential decision process and introduce the RL framework,where the overall learning target is to find the best global layout and the most adequate local details.

Fig. 1 depicts the pipeline and how the overall learning for collage generation is designed from state space, action space and reward function. The state and action space consider the global layout and local details, and the reward function is designed to consider subjective and objective factors. The overall optimization goal is to maximize the accumulated reward of the trajectory generated by the agent's policy.

1) State Space:

The agent keeps interacting with the environment by observing from the current state $s \in S$ of the environment and performs actions $a \in A$ according to the policy $\pi(a|s)$.

The observation o_t includes the current aesthetic feature extracted from the collage concatenated with the initial feature. The state $s_t = \{o_0, o_1, ..., o_t\}$ includes all the past observations. Taking advantage of historical experience, a long short-term memory (LSTM) unit is added in the agent to assist in making an improved policy.

TABLE II: The action space design considers global layout and local details. The global layout actions aim to improve the composition quality and the local detail actions assists to achieve the fine-grained collocation of sub-images.

Category	Attribute	Operation
global layout (C1)	layout	switch image-pair
	termination	-
local details (C2)	x-relative-position	-15/-5/0/+5 (pixel)
	y-relative-position	-15/-5/0/+5 (pixel)
	layer	top [1]/bottom [1]/ -
	orientation angle	-0.5/0/+0.5 (°)

[1] The "top" (or "bottom") operation puts an image on the top (or bottom) layer to highlight (or hide) it.

2) Action Space:

Inspired by the quick initialization in [11] that optimizes the collage from the layout to the details, we customize the action space for collage generation into two categories to adjust the global layout and local details with different attributes. The detailed attributes and operations for each category are listed in Table II.

Global Layout. The global layout exerts substantial effects on composition quality and is essential for a visually satisfying collage. The main factors that influence composition quality are balance and emphasis [32].

Inspired by long-distance image dragging in manual collages, our agent switches one image pair at each step before reaching the maximum step. The switch action can be interpreted as changing the order of input images. It affects the global layout through optimizations of image adjacency collocations and is executed in multiple consecutive steps to adjust the positions of images of high importance (e.g., group photos) in a collage. Given that the agent is expected to decide the best layout when the score no longer increases, the termination action is designed as a trigger to stop the transformation process and output the current layout.

Local details. For detail adjustments, the agent operates on each individual image. Inspired from the state variables defined on the image set in [11], the detail adjustment in this work is designed with three attributes: spatial position, placement order, and orientation angles. Each attribute helps capture the fine-grained pleasantness expressed by users. Specifically, spatial position adjustment introduces the overlay style and increases informativeness. The layer index determines the placement order of the input image collection. The rotation for each image increases the orientation diversity and satisfies natural preferences.

Directly operating on the coordinates of each individual image is inefficient due to the complex search space of spatial search space. Alternatively, we formulate the learning of absolute spatial positions of separate images as relative displacement of adjacent images. The overlay area between images is initialized as zero and altered progressively in one episode. The agent adjusts the position of each image by refer-

ring to the neighboring image. With regard to the placement order, the layer operation is designed to place an image on the top or bottom layer to highlight the essential content and hide the irrelevant area. Then, the rotation operation rotates each image by 0.5° clockwise or counterclockwise at each step to improve the visual impression of the collage results.

3) AutoCrop Module:

The AutoCrop Module adjusts a collage with irregular shapes to the aspect-ratio-specified canvas during each episode. It incorporates the aspect ratio information into the environment and gives feedback to the agent at each step. Specifically, after the agent adjusts the collage at each step, multiple candidate views are cropped from the current collage, resulting in consistency with the canvas. Then, view selection is completed in the evaluation network, and the cropped collage with the highest score is sent to the next step. As a result, the agent is encouraged to choose actions that progressively avoid losing salient information while suppressing the blank space on the canvas.

4) Reward Design:

The reward function describes the preference for the current state, and the overall target is to find the most visually pleasing collage result. To achieve this target, the reward function considers subjective and objective scores to provide the agent an incentive to optimize the collage quality at each step.

Considering that the major challenge in analyzing photo collages is structural complexity, we assess the subjective quality with composition quality. Specifically, we use the aesthetic evaluation network and propose the utilization of the aesthetic proposal number satisfying the aesthetic selection standard as representation for the subjective score of collage C, which is denoted as $s_a(C)$. Intuitively, collage results with a large aesthetic box number indicate high aesthetic quality. For the objective score, because salience is implicitly optimized by the aesthetic network, we calculate the blank area s_b on the canvas. Overall, the evaluation score for a collage is computed as:

$$s(C_t) = \lambda_a s_a(C_t) - \lambda_b s_b(C_t) \tag{5}$$

where the λ_a and λ_b stand for the weights of the aesthetic and blank loss items, respectively. After each step, the difference in the aesthetic score between the updated collage C_{t+1} and the previous collage C_t is used to calculate the reward for the current policy. To increase the aesthetic quality while suppressing the blank space, the agent is granted a positive reward if the score increases and a negative reward otherwise.

$$r_t^{'}(C_t) = s(C_{t+1}) - s(C_t) \tag{6}$$

Finally, the agent is facilitated with the greedy strategy to avoid redundant actions and speed up the generation progress because the reinforcement reward scheme indirectly treats the number of steps as a potential cost.

$$r_t(C_t) = r_t^{'}(C_t) - 0.01 * (t+1) \tag{7}$$

5) Training Algorithm:

We adopt the A2C algorithm as our RL framework to train the policy of collage generation. The A2C includes two sub-networks. The policy network θ_p outputs the probability distribution over the designed action space, each corresponding to the action operating on the collage according to the policy $\pi(a^{(t)}|s^{(t)})$. The value network outputs $V(s_t;\theta_v)$, which predicts the expected accumulated reward R_t at step t.. Both networks share the backbone to reduce the parameters. The global reward R_t is estimated as $r^t + \gamma V(s^t)$ and the overall optimization target during training is described as follows:

$$L_{\theta_p} = -\log \pi(a \mid s^{(t)})(R^{(t)} - V(s^{(t)})) + H(\pi(s^{(t)})) \tag{8}$$

$$L_{\theta_v} = (R^{(t)} - V(s^{(t)}))^2 \tag{9}$$

where the optimization goal for the policy network is to maximize the advantage function computed as $R^{(t)} - V(s^{(t)})$ and the entropy $H(\pi(s_t;\theta))$ of policy output. The entropy in the optimization objective aims to increase the diversity of actions, which encourages the agent to learn flexible policies.

IV. Experiment

In this section, we first explain the experiment settings (Section IV-A). Second, we introduce the aesthetic evaluation protocol and three other basic criteria exploited in the most previous state-of-the-art work (Section IV-B). Third, quantitative and qualitative evaluations are conducted in Sections IV-C and IV-D, respectively.We also perform ablation studies on the proposed modules in Section IV-E. Lastly, we investigate the relationship between subjective preferences and the proposed aesthetic evaluation method through user studies (Section IV-F).

A. Experiment Settings

To validate the scalabilities, we evaluate the proposed model on both videos and image sets.

Dataset. For video evaluation, we collect 54 videos from the Hollywood2 movie dataset [33], which is composed of 12 classes of human actions and 10 classes of scenes, and three movies from the MPII movie description dataset (LSMDC3) [34]. The three movies are "The Queen," "Up in the Air," and "Pride and Prejudice." The dataset provides comprehensive realistic movie scenes with challenging settings.

For the Hollywood2 dataset, we extract the key frames and resize them to 540 × 900. Each video consists of 40 randomly sampled sets of images with different numbers for composing the final collage. Meanwhile, the LSMDC3 dataset provides image collections extracted from sequential time clips. To make meaningful collages and meet the needs of real-world scenarios, the image series in LSMDC3 are selected from the same context scene, and duplicated frames are removed within the same period. A total of 400 and 600 image sets are included in the two test sets.

For the image set evaluation, we collect 1000 personal vacation images from the Pexels website [1]. The images are selected based on popularity and have rich content. The images are preprocessed to 600 × 900, and 400 image sets are included in the evaluation dataset.

[1]https://www.pexels.com

Implementation Details. We train the model with $\sim$ 2,000 image sets included in the Hollywood2 dataset, from which our agent learns robust policies from videos with diverse quality. The model is evaluated on two video test sets and finetuned on image sets. Our aesthetic evaluation is outputted by the View Proposal Network pretrained on the CPC dataset [18], which is an aesthetic ranker with advanced image aesthetic evaluation accuracy. The evaluation and deep aesthetic network share the same feature-extracting unit to stabilize the training process.

During training, the max epoch is set to 50. For stability reasons, a signal function for reward is used the first 20 epochs and removed for the remaining 30 epochs. The max step is set to 12. We set 32 for batch size and use the Adam optimizer [35]. The learning rate and weight decay are set to $1e$–3 and $1e$–5, respectively.

B. Evaluation Metrics

To assess the quality of the collage results comprehensively, we evaluate the results in terms of informativeness, canvas coverage, visual balance, and aesthetics. The first three metrics are the most common criteria used in state-of-the-art photo collage methods. The aesthetic metric is developed to describe the overall quality of the collage results.

Informativeness. Considering that the salient object has a great influence on visual perception, we use salience preservation as the metric for evaluating the informativeness of the collage results. We obtain a salience map for each image in accordance with [36] and compute the informativeness as

$$M_{\text{info}} = \frac{1}{T}\sum_{i=1}^{T}\frac{\text{Sal}\left(\text{Vis}\left(M_i\right)\right)}{\text{Sal}\left(M_i\right)} \tag{10}$$

where T denotes the number of input images, M_i represents the salience map of the i-th image, $Vis(\cdot)$ computes the visible parts, and $Sal(\cdot)$ computes the sum of non-black pixel numbers.

Canvas Coverage. We use the blank area ratio to represent the canvas coverage. The blank area is the space in the canvas that is not covered by any photo. The coverage ratio is computed as

$$M_{\text{cov}} = \frac{\bigcup_{i=1}^{T} s\left(I_i\right)}{s(C)} \tag{11}$$

where I_i represents the i-th input image.

TABLE III: Quantitative evaluation of the aesthetic quality for photo collages with different methods generated on the Hollywood2 and LSMDC3 dataset.

Methods	**Aesthetic Score**	
	Hollywood2	LSMDC3
Instagram Layout [37]	79.83	109.06
Shape Collage [38]	88.22	114.68
Circle Packing Collage [7]	63.12	106.84
Picture Collage [14]	83.25	112.36
AutoCollage [10]	103.67	132.65
Ours (w/ AutoCrop)	**104.81**	**135.13**
Ours (w/o AutoCrop)	**110.62**	**138.94**

Visual Balance is used to measure the composition quality and is an essential factor during the creation of visual displays.

Fig. 2: Visual comparisons of the impact of the proposed actions on the global layout and local details of the generated collage. (a) The baseline method is the quick initialization result from Picture Collage which arranges the layout mainly considering salience energy. (b) The C1 actions transform the global layout and get the aesthetic layout augmented with rule of center. (c) The C2 actions adjust the local details of individual images and assist to highlight the relevant aesthetic frames, after which the agent generates the aspect ratio specified result (eg. "3:4") as the output. (d) The blending style could be optionally added to the collage result along the boundaries for the purpose of seamless transition between adjacent images.

We exploit the metric Mea introduced in [5] to measure the effectiveness of the layout results.

Aesthetics. The aesthetic score is designed to describe the overall quality of the collage results. It is computed as

$$M_{\text{aes}} = \sum_{i}^{N} s\left(P_i\right)\cdot f\left(P_i\right)\cdot \delta\left(\frac{s\left(P_i\right)}{s(C)} > \eta\right) \tag{12}$$

where $s(P_i)$ and $f(P_i)$ represent the area and score of the i-th patch, respectively, and N denotes the proposal box number satisfying the selection criteria and collage results. A large aesthetic box number indicates high aesthetic quality. η is set to 60% by default in practice.

The evaluation metrics consider the size and quality of different local regions, and the intuition behind the metric here is to assess the global quality of a holistic collage with

accumulated local composition quality of sub-collage-parts. With Equation 12, we assumed that a high-quality collage accumulates a high aesthetic score over the local parts by means of numerous balanced compositions, less occlusion along boundaries, or fewer blending artifacts.

C. Quantitative Evaluation

To assess the effectiveness of the proposed model, we evaluate different methods quantitatively. We compare the proposed model's aesthetic score with those of several competing methods in the first section. We examine the effect of our action space design and evaluate the global layout and local details quality of a collage in the second section. We evaluate the proposed method by using three basic criteria in the last section.

1) Evaluation of Aesthetic Quality:

We conduct quantitative comparisons with other existing methods to evaluate the effectiveness of the proposed network.

The competing methods are (a) AutoCollage [10], which creates a collage of representative elements from an image set and develops a sequence of optimization steps for collage generation; (b) Circle Packing Collage [7], which partitions a canvas by using the importance of regions of interest from input images; (c) Picture Collage [11], which addresses the photo collage issue with handcrafted energy terms and generates collages through quick initialization and Markov chain Monte Carlo optimization; (d) Instagram Layout[37], which is an app developed by Instagram that combines multiple photos into one single image with predefined templates; and (e) Shape Collage[38],which is an automatic photo collage maker that allows users to make shapes or blend collages of family photos in a harmonious way via many flexible templates.

To generate results from the same test set, we run a simulation click program on a compiled software to automatically generate the results of AutoCollage, Circle Packing Collage, Instagram Layout, and Shape Collage. Given that AutoCollage Touch 2009 has an input number limit, it can only generate collages with more than six input images. Picture Collage also achieves competitive results, so we reimplement its quick initialization process to perform comparisons.

As shown in Tables III and V, our method achieves consistently higher aesthetic scores than the competing methods in video and image datasets.

2) Evaluation of Action Space Design:

To evaluate the effectiveness of the proposed action space design in Table II, we examine the effect of layout adjustment (C1) and local detail optimization (C2) on the improvement of collage aesthetic quality. Considering that Picture Collage also generates collages with spatial coordinates, rotation angles, and layer indices and performs optimization through handcrafted energy terms, we reimplement the quick initialization of Picture Collage as the baseline method and quantitatively compare the quality of the intermediate layout and the final results of both methods. Then, we evaluate them based on their proposal number and aesthetic score. The visual comparison of different action sets is illustrated in Figure 2.

For global layout adjustment, we use the layout initialization from Picture Collage as the initialization for our network. We compare the layout results of our proposed network with those of the baseline method under multiple different-numbered inputs. As shown in Table III, the agent learns to improve aesthetic quality through image-pair-switch operations and helps create the global layout with increased proposal views, thereby increasing the aesthetic score.

For local detail refinement, both methods are initialized with the same layout for fairness. Compared with the baseline method that pays special attention to salience constraints, our method receives feedback from subjective and objective factors. Table IV shows that our results achieve better improvement compared with salience-based optimizations.

3) Evaluation of Basic Criteria:

We evaluate the collage results via three basic criteria exploited in most state-of-the-art works; these three are informativeness, canvas coverage, and visual balance. To assess the composition quality of our layout results, we compare our results against those of a random layout, the layout that optimizes salience balance in rows and columns [11], and the layout that optimizes balance [5]. Given that the original model and weights of [5] are unattainable, we replace the aesthetic reward with Mea in [5] and report the composition quality results in Table VI. The results demonstrate that the aesthetic reward brings visual balance.

To assess the detail adjustment quality under different scenarios, we conduct experiments on the collage results with four mainstream aspect ratios. The results are summarized in Table VII and reveal the effectiveness of the detail adjustment actions.

D. Visual Comparisons with Existing Methods

We qualitatively compare our methods against several competitors with three typical collage styles(described in Section IV-C1), namely, blending, overlay and template style. The visual comparisons are displayed in Fig. 4.

The blending style-based methods are AutoCollage [11] and Circle Packing Collage [7]. As shown in Figure 4, AutoCollage has multiple artifacts and loses salient information along

TABLE IV: Aesthetic evaluation on the global layout and local details compared with the baseline method on the Hollywood2 dataset. The second row implies the number of input images, with fixed or unfixed number at one time. The C1 and C2 denote the two categories of proposed action space in Table II.

Methods	Evaluation	Proposal Number			Aesthetic Score		
	Input No.	6	8	15	6	8	15
Baseline		10.9	10.16	11.34	87.18	81.25	85.5
Ours(w/ C1)		12.31	11.49	12.53	97.51	93.81	93.72
Ours(w/ C2)		13.08	12.06	13.15	101.71	95.02	98.36

TABLE V: Quantitative evaluation of the aesthetic quality for photo collages with different methods on the image dataset.

Methods	Aesthetic Score
Instagram Layout [37]	101.02
Shape Collage [38]	106.62
Circle Packing Collage [7]	98.43
Picture Collage [14]	103.88
AutoCollage [10]	130.78
Ours (w/ AutoCrop)	**132.97**
Ours (w/o AutoCrop)	**137.09**

TABLE VI: Quantitative evaluation of the composition quality for the first stage of collage results on the LSMDC3 dataset.

Methods	Mea
Random Layout	23.06
Picture Collage [14]	12.13
Grid Collage [5]	9.02
Ours(w/ aes)	**9.32**
Ours(w/ aes+mea)	**9.18**

the boundaries. It also fails to highlight the visual focus when dealing with complex scenes. Circle Packing Collage loses most of the salient information and generates confusing boundaries in most cases. Picture Collage [11] introduces the overlay style to avoid blending artifacts, but it is based on salience energy optimization and is unable to generate an aesthetic photo collage with good composition quality. Instagram Layout[37] and Shape Collage [38] are template style-based methods. The former generates collages based on the input order, and the latter randomly generates photo collages with rich grid templates. However, both methods rely heavily on user interaction in practice.

Compared with the competing models, our method achieves better composition quality while preserving the local details. The irrelevant area is greatly reduced via occlusion, and the salient object is well-preserved. Moreover, the results of our proposed method have clear boundaries and avoid artifacts.

We visualize the results of the global layout and the optimized results of the generated collage in Figure 2. With the proper layout (center rule) and detail adjustment, the salient content is highlighted in the center area. The idea behind the proposed method is similar to the process of humans making collages, thereby making the automatic generation process interpretable. To further improve scalability, the blending style can be added to the collage result along the boundaries for the purpose of seamless transition between adjacent images.

TABLE VII: Quantitative evaluation of the informativeness and canvas coverage of detail adjustment results on the LSMDC3 dataset.

Metric	Aspect Ratio			
	3:4	**1:1**	**4:3**	**16:9**
M_{info}	0.80	0.76	0.79	0.83
M_{cov} **(%)**	0.03	0.03	0.04	0.07

E. Ablation Study

To prove the effectiveness of the proposed components, we perform ablation experiments to prove the functions of the attention fusion and AutoCrop modules and verify the reasonableness of the evaluation metric.

1) Attention Fusion Module:

We examine the influence of the attention fusion module on the aesthetic quality and centrality [12] of the generated collage results and investigate the effect of the attention mechanism.

Fig. 3: Results of the proposed model from LSMDC3 and the collected image dataset on four mainstream aspect ratio specified canvases(i.e., ”1:1”, ”16:9”, ”3:4”, ”4:3”).

Fig. 4: Visual comparisons on the LSMDC3 dataset with competing methods. Artifacts along the boundaries of images are annotated with red boxes. In both scenarios, the AutoCollage[10] presents multiple artifacts along the blending borderline, the Circle Packing Collage[7] loses most salient information, the Picture Collage[14] loses salient information on the canvas border and struggles at highlighting the visual focus and the Instagram Layout[37] and the Shape Collage[38] suffers from content loss due to inappropriate cropping along the border. By contrast, our method exhibits the strengths with balanced global layout and preserved local details. Moreover, the salient content is highlighted in the center area while the irrelevant area is decreased with the occlusion.

TABLE VIII: Quantitative evaluation of the centrality for collage results on the LSMDC3 dataset.

Methods	Input Number			
	6	8	9	12
PC [11]	0.514	0.455	0.389	0.486
Ours	0.618	0.584	0.765	0.604

The attention layer shifting its focus to the center of the collage is an implicit utilization of central rules and can thus enhance the subjective quality of the collage results. After we remove the attention layer, the aesthetic score decreases by 4%.

Additionally, we conduct experiments on the centrality of the collage results. The centrality metric was introduced in [12] and is computed as

$$M_{centrality} = 1 - \frac{\|\mathbf{c}_1 - \mathbf{c}_0\|_2}{0.5 * \text{diag}(\mathcal{C})} \tag{13}$$

where c_0 represents the centroid of the canvas, c_1 represents the centroid of the visible part of the image on the top layer, and $diag(\cdot)$ represents the diagonal length of the canvas. The idea behind centrality is that the most important photos are placed at the top of the less important ones to minimize the risk of being severely occluded. The results in Table VIII prove the effectiveness of the attention fusion module.

2) AutoCrop Module:

We investigate the effect of the AutoCrop module on mainstream aspect-ratio-specified canvases to improve the effectiveness in realistic applications.

The AutoCrop module adapts a generated collage with irregular shapes to an aspect-ratio-specified canvas. Given that the AutoCrop module is based on sliding windows powered by the aesthetic network, it can help choose the best view from a raw collage while maximizing the area of concern on the aspect-ratio-specified canvas.

To prove the effectiveness of the AutoCrop module, we perform a test on four mainstream aspect ratios (i.e., 1:1, 16:9, 4:3, and 3:4) and test the module's capability both evaluation datasets. The collage results are shown in Figure 3. Under a compact aspect ratio, the agent learns to increase the occlusion in order to avoid cropping out salient information; in the opposite case, the agent learns to decrease the overlay area in order to minimize the blank space.

The results in Tables III and V also show that the content of concern is considerably preserved without being cropped by canvas borders and demonstrate the module's effectiveness in video and image datasets.

3) Evaluation Metric:

We explore the plausibility of the evaluation metric proposed in Equation 12. Specifically, we modify η in the fusion process and change it to 50% and 40%, respectively. Although lowering the threshold promises more aesthetic proposals, the increasing number of proposal views brings noisy signals and

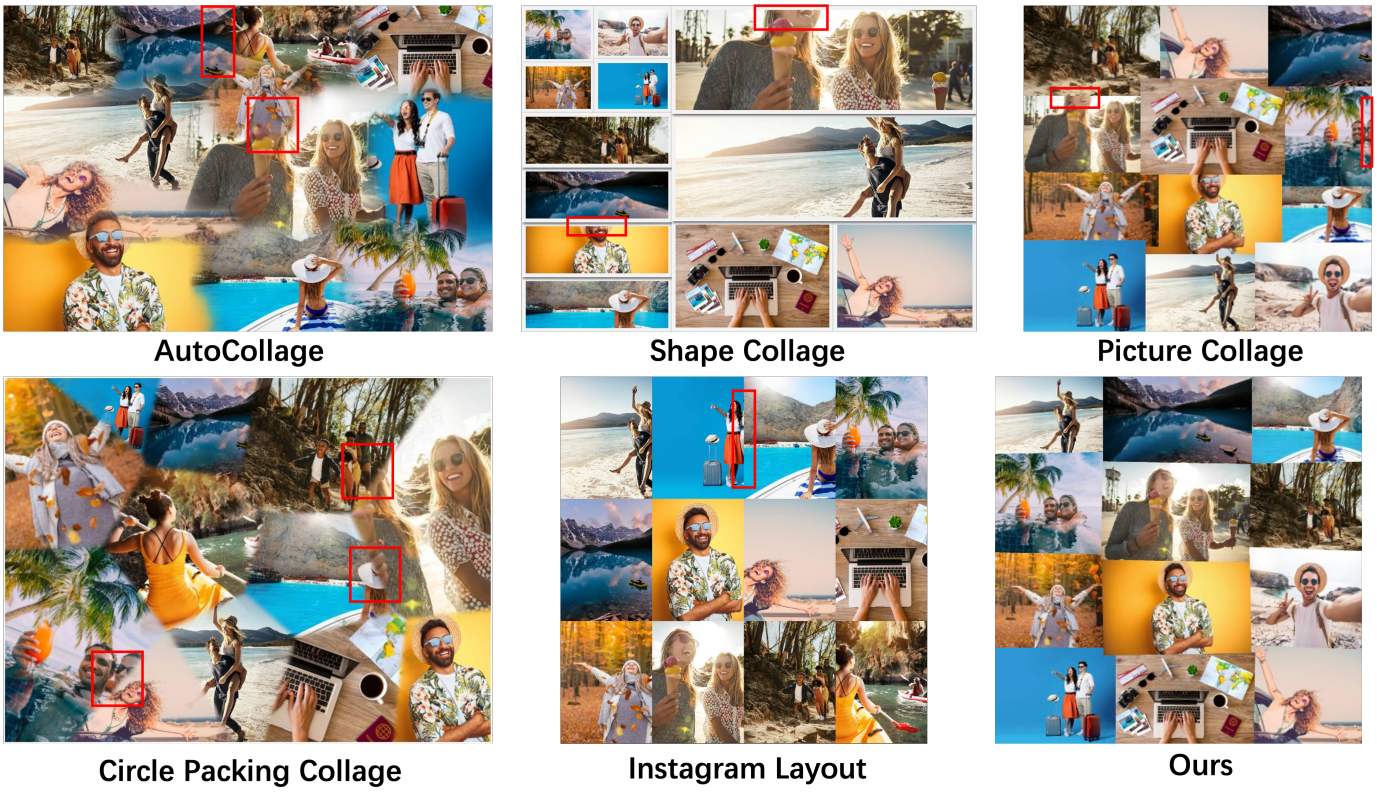


Fig. 5: Visual comparisons on the image dataset with competing methods, including AutoCollage[10], Shape Collage[38], Picture Collage[14], Instagram Layout[37], Circle Packing Collage[7]. Artifacts along the boundaries of images are annotated with red boxes.

causes oscillation to the training process and the performance decreases by 1.5% and 4.85%. Also, increasing the threshold to 70% reduces the structural information and leads to a 1.25% performance drop.

4) Limitations:

One limitation of our method is the aspect ratio of the input sequence. Since the discussion was mainly focused on the deep aesthetic feature representation and two-phase generation for aspect-ratio-specified collages, the input images are preprocessed to the same shape for the network. Another limitation is the input size. The input sizes are configured as not more than 15 due to the local optima problem of deep RL. In addition, user's interaction is not designed as part of the input because the goal of our method is to automatically generate aesthetic results and simplify the process.

F. User Study

Apart from quantitative collage evaluations, we conduct two user studies to evaluate the effectiveness of our algorithm. In the first experiment, we evaluate the quality of different collage methods and investigate the correlation with the aesthetic score. In the second one, we investigate the aesthetic quality of the proposed method under different criteria. To evaluate the results on different types of inputs, we conduct evaluations for image and video datasets separately in each part.

Evaluation against different methods. To assess different collage methods subjectively, we conduct a user study with six different methods by using questionnaires. We prepare 24 groups of photo collages from the video dataset and 24 groups of photo collages from the image dataset by using six different methods. Each group consists of different methods using the same input. The order in each group is shuffled to guarantee fairness. Twenty users not involved in the study are invited as participants. The observers do not have special experience in making collages, so we provide them good results demonstrated in related studies as positive cases before the experiment. For each group, the participants are asked to watch for not less than 20 seconds and then rate the collages from beautiful (6) to unattractive (1) on the basis of personal aesthetic standards. The criteria include no occlusion of significant objects, few blank areas on the canvas, balanced composition quality, and natural display of images. Figure 6 lists the average scores of each method. The proposed method has high evaluation scores on both video and image datasets, which prove its effectiveness.

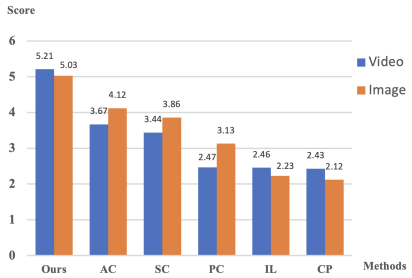


Fig. 6: User studies on different photo collage methods(from left to right, our method, AutoCollage[10], ShapeCollage[38], Picture Collage[14], Instagram Layout[37], Circle Packing Collage[7]). User studies report results in line with the aesthetic score in Tables III and V, which is a further proof of the effectiveness of proposed evaluation method.

To evaluate the correlation between the proposed aesthetic score and the participants' opinions, we select the top-voted competitor from three kinds of collages with different style (i.e., blending, overlay and grid style).

TABLE IX: The correlation of the metric between users' opinions and the aesthetic score on the top-voted methods.

Type	Metric	PC	SHP	AC	Ours	PRCC
Video	M_{aes}	83.25	88.22	103.67	110.62	0.9058
	M_{user}	2.47	3.44	3.67	5.21	
Image	M_{aes}	103.88	106.62	130.78	137.09	0.8814
	M_{user}	3.13	3.86	4.12	5.03	

We compute the Pearson correlation coefficient (PRCC) of the four methods with different style groups as follows:

$$\text{PRCC} = \frac{\text{cov}(M_{aes}, M_{user})}{\sigma_{M_{aes}} \sigma_{M_{user}}} \tag{14}$$

where the PRCC result of 0.9058 and 0.8814 proves the high correlation between the proposed metric and user's opinions.

Evaluation under different criteria. To perform a comprehensive assessment of the collages, we ask users for their opinions under different criteria. We prepare 15 groups of collages generated from the video dataset and 15 groups of collages generated from the image dataset used in the experiment and invite another 20 observers not involved in this work as participants. For each collage, the raters are required to watch for 10 seconds then answer the following questions by using a scale of 1 (definitely no) to 5 (definitely yes). To make the results consistent, we provide some reference standards in the parentheses as deduction points.

- Q1: Do you think the collages have a natural display of images and overall excellent quality?
- Q2: Do you think the collages convey information sufficiently (significant objects are not occluded nor cropped along the border)?
- Q3: Do you think the images cover the canvas sufficiently (no holes or inappropriate blank areas are present in the images or along the border)?
- Q4: Do you think the composition quality is well-balanced and harmonious (balanced layout and proper image arrangement)?

The results for each type of input have an average of 4.4 and 4.375, indicating that our method creates visually pleasing collages.

TABLE X: Statistics of users' opinions on photo collages generated on image and video datasets under different criteria.

Data Type	Q1	Q2	Q3	Q4
Video	4.4	4.3	4.5	4.4
Image	4.5	4.2	4.6	4.2

V. Conclusion

In this study, a novel pipeline for automatic photo collage generation is proposed. Inspired by manual collages, the collage generation task is decomposed into interpretable steps and modeled as an RL process. The attention fusion module embedded in the deep aesthetic network overcomes the lack of training data and provides a comprehensive feature representation for the photo collage. Moreover, the AutoCrop module is proposed to inherently generate an aspect-ratio-specified collage, thus making the application scenario highly flexible. The experiments on video and image datasets demonstrate the superiority of the proposed model, and the user studies prove the effectiveness of the subjective evaluation and our method. In the future, we will consider exploring rich action spaces to improve informativeness and introduce interactive design for fulfilling personalized opinions. Experiments on video and image datasets demonstrate the superiority of the proposed model, and the user studies further prove the effectiveness of the subjective evaluation and our method.

References

[1] C. B. Atkins, "Blocked recursive image composition," in *ACM International Conference on Multimedia*, 2008, pp. 821–824.

[2] Z. Wu and K. Aizawa, "Picwall: Photo collage on-the-fly," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2013, pp. 1–10.

[3] Y. Liang, X. Wang, S.-H. Zhang, S.-M. Hu, and S. Liu, "Photorecomposer: Interactive photo recomposition by cropping," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 10, pp. 2728–2742, 2017.

[4] X. Pan, F. Tang, W. Dong, C. Ma, Y. Meng, F. Huang, T.-Y. Lee, and C. Xu, "Content-based visual summarization for image collections," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 4, pp. 2298–2312, 2019.

[5] Y. Song, F. Tang, W. Dong, F. Huang, T.-Y. Lee, and C. Xu, "Balance-aware grid collage for small image collections," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2021.

[6] S. Goferman, A. Tal, and L. Zelnik-Manor, "Puzzle-like collage," in *Computer Graphics Forum*, vol. 29, no. 2. Wiley Online Library, 2010, pp. 459–468.

[7] Z. Yu, L. Lu, Y. Guo, R. Fan, M. Liu, and W. Wang, "Content-aware photo collage using circle packing," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 2, pp. 182–195, 2013.

[8] L. Liu, H. Zhang, G. Jing, Y. Guo, Z. Chen, and W. Wang, "Correlation-preserving photo collage," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 6, pp. 1956–1968, 2017.

[9] L. Tan, Y. Song, S. Liu, and L. Xie, "Imagehive: Interactive content-aware image summarization," *IEEE computer graphics and applications*, vol. 32, no. 1, pp. 46–55, 2011.

[10] C. Rother, L. Bordeaux, Y. Hamadi, and A. Blake, "Autocollage," *ACM Transactions on Graphics*, vol. 25, no. 3, pp. 847–852, 2006.

[11] T. Liu, J. Wang, J. Sun, N. Zheng, X. Tang, and H.-Y. Shum, "Picture collage," *IEEE Transactions on Multimedia*, vol. 11, no. 7, pp. 1225–1239, 2009.

[12] S. Bianco and G. Ciocca, "User preferences modeling and learning for pleasing photo collage generation," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 12, no. 1, pp. 1–23, 2015.

[13] C. Rother, S. Kumar, V. Kolmogorov, and A. Blake, "Digital tapestry [automatic image synthesis]," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1. IEEE, 2005, pp. 589–596.

[14] J. Wang, L. Quan, J. Sun, X. Tang, and H.-Y. Shum, "Picture collage," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1. IEEE, 2006, pp. 347–354.

[15] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "Rapid: Rating pictorial aesthetics using deep learning," in *ACM International Conference on Multimedia*, 2014, pp. 457–466.

[16] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang, "Deep multi-patch aggregation network for image style, aesthetics, and quality estimation," in *IEEE International Conference on Computer Vision*, 2015, pp. 990–998.

[17] K. Sheng, W. Dong, C. Ma, X. Mei, F. Huang, and B.-G. Hu, "Attention-based multi-patch aggregation for image aesthetic assessment," in *ACM International Conference on Multimedia*, 2018, pp. 879–886.

[18] Z. Wei, J. Zhang, X. Shen, Z. Lin, R. Mech, M. Hoai, and D. Samaras, "Good view hunting: Learning photo composition from dense view pairs," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5437–5446.

[19] Y. Wei, Y. Matsushita, and Y. Yang, "Efficient optimization of photo collage," *Microsoft Research, Redmond, WA, USA, MSRTR-2009-59*, 2009.

[20] Y. Kao, R. He, and K. Huang, "Deep aesthetic quality assessment with semantic information," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1482–1495, 2017.

[21] H. Zeng, Z. Cao, L. Zhang, and A. C. Bovik, "A unified probabilistic formulation of image aesthetic assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 1548–1561, 2019.

[22] W. Wang and J. Shen, "Deep cropping via attention box prediction and aesthetics assessment," in *IEEE International Conference on Computer Vision*, 2017, pp. 2186–2194.

[23] H. Zeng, L. Li, Z. Cao, and L. Zhang, "Reliable and efficient image cropping: A grid anchor based approach," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5949–5957.

[24] Y.-L. Chen, J. Klopp, M. Sun, S.-Y. Chien, and K.-L. Ma, "Learning to compose with professional photographs on the web," in *ACM International Conference on Multimedia*, 2017, pp. 37–45.

[25] S. Ma, J. Liu, and C. Wen Chen, "A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4535–4544.

[26] D. Li, H. Wu, J. Zhang, and K. Huang, "A2-rl: Aesthetics aware reinforcement learning for image cropping," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8193–8201.

[27] Li Debang, Wu Huikai, Zhang Junge and Huang, Kaiqi, "Fast a3rl: Aesthetics-aware adversarial reinforcement learning for image cropping," *IEEE Transactions on Image Processing*, vol. 28, no. 10, pp. 5105–5120, 2019.

[28] J. Park, J.-Y. Lee, D. Yoo, and I. So Kweon, "Distort-and-recover: Color enhancement using deep reinforcement learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5928–5936.

[29] K. Yu, C. Dong, L. Lin, and C. Change Loy, "Crafting a toolchain for image restoration by deep reinforcement learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2443–2452.

[30] B. Zhong, B. Bai, J. Li, Y. Zhang, and Y. Fu, "Hierarchical tracking by reinforcement learning-based searching and coarse-to-fine verifying," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2331–2341, 2018.

[31] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European Conference on Computer Vision*. Springer, 2016, pp. 21–37.

[32] J. Geigel and A. C. Loui, "Automatic page layout using genetic algorithms for electronic albuming," in *Internet Imaging II*, vol. 4311. International Society for Optics and Photonics, 2000, pp. 79–90.

[33] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 2929–2936.

[34] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville, and B. Schiele, "Movie description," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 94–120, 2017.

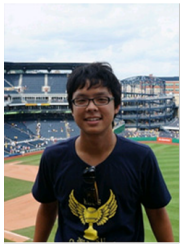
[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Arxiv Preprint arXiv:1412.6980*, 2014.

[36] G. Lee, Y.-W. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 660–668.

[37] Instagram Inc. (2015) Instagram layout. [Online]. Available: https://play.google.com/store/apps/details?id=com.instagram.layout

[38] V. Cheung. (2013) Shape collage. [Online]. Available: http://www.shapecollage.com/

Mingrui Zhang is currently a postgraduate student at the Institute of Computer Graphics and Computer Aided Design, School of Software, Tsinghua University. He received B.S. degree (Hons.) in computer science from Beijing University of Posts and Telecommunications in 2020. His research interests include computer vision and image processing.



Mading Li is currently an Algorithm Engineer with Video Technology Team, Kuaishou, Beijing, China. His research interests focuses on image/video quality evaluation and smart video editing. He received the B.S. degree in computer science from Peking University, Beijing, China, in 2013, and the Ph.D. degree from the Institute of Computer Scienceand Technology, Peking University, in 2018. He was a Visiting Scholar with McMaster University, Hamilton, ON, Canada, from 2016 to 2017.

Jiahao Yu received the B.S. degree from the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications. He is currently working toward the M.S. degree in the School of Software, Tsinghua University. His current research interests include computer vision and machine learning.

Li Chen received the PhD degree in computer graphics from Zhejiang University, China, in 1996. She is currently a professor with the Institute of Computer Graphics and Computer Aided Design, School of Software, Tsinghua University, China. Her research interests include data visualization, image processing, and computer graphics.