

Aesthetic Photo Collage with Deep Reinforcement Learning

Mingrui Zhang, Mading Li, Li Chen, Jiahao Yu

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 collage for the first time. 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 two movie datasets and our results outperform others in aesthetic quality evaluation. A further user study is 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 steep surge of images and videos has lead 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 personal photo summarization. Yet the scalability and flexibility also make it a challenging task to generate photo collage with high aesthetic quality.

In past decades, many works have been published addressing this issue. One method for solving this problem is based on canvas partitioning, which typically involves circle packing algorithm [1] or feature embedding [2] to partition the canvas into separate disjointed areas. Another method employs customized features [3], [4] to assess the quality of a collage and minimize the energy term through complex optimizations.

However, such handcrafted-based methods cannot provide adequate collage representation and generate high-quality collages of general scenes. To overcome the limitations of handcrafted features, deep learning can provide comprehensive feature representation. However, owing to the subjectivity and complexity of collage tasks, training data is lacking and

unsuitable for supervised learning. Moreover, photo collage is a multistep task, which increases the infeasibility of directly applying deep learning.

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

Since manual annotations for photo collage require highly skilled designers, and photo collages are complex and need a substantial amount of training data; thus, the high cost involved makes constructing a training dataset for photo collages unrealistic. Consequently, we pretrain our aesthetic network on a large scale image aesthetic dataset (i.e., Comparative Photo Composition (CPC)) [5] for general aesthetic feature extraction and propose an attention fusion module for collage feature extraction. The attention fusion module is designed to extract the complex structural features of a photo collage, such as the composition and collocation of different subimages. Specifically, for better adaptation in photo collage, the attention fusion module adopts multi-patch information with an attention mechanism to effectively represent the complex features of a collage, inspired by image aesthetic evaluation[6], [7], [8].

With the aesthetic and structural feature representation from deep aesthetic network, we formulate collage generation as a sequential decision process and present an improved RL framework. Specifically, we design a policy network to manipulate the global layout and local detail properties of individual images, including the orientation angle, relative spatial position, and placement order. In each step, the policy network makes improvements and generates an aspect ratio-specified collage, while the value network assists in stable and precise policy making. The reward design considers subjective and objective factors to instruct the agent to generate collage results with a balanced composition and less blank spaces. The policy making training is adapted to the advantage actor-critic (A2C) algorithm.

In summary, our main contributions are as follows:

- We propose a novel pipeline for automatic photo collage generation. We decompose the collage generation into interpretable steps and model it as an reinforcement learning process, which to our knowledge, is the first work for directly applying deep learning in automatic photo collage.

- We develop the deep aesthetic network for general aesthetic feature extraction and propose an attention fusion module for structural collage feature representation. To overcome the lack of collage dataset, we pretrain the main aesthetic network on large scale image aesthetic dataset (CPC).
- We evaluate our model against several competing methods on the Hollywood2 and LSMDC3 movie dataset. Our model outperforms the other methods in aesthetic quality evaluation. To demonstrate the effectiveness of our model and the subjective visual quality, we conduct a user study.

II. RELATED WORKS

Photo collage aims to create a visually appealing summary by arranging multiple images on a given canvas. Previous works on photo collages mainly fall into three categories. (a.) Region partitioning-based methods [1], [2] involve a circle packing algorithm [6] or feature embedding [2] to partition the canvas into separate disjointed areas. (b.) Content preserving collages [9], [10] rely on tree-based page division to recursively split a canvas but ignore the image content. (c.) Customized energy terms optimizations-based methods [3], [4], [11] cast photo collage as an optimization problem with well-defined objective functions. Compared with traditional methods which focus mainly on optimizing regions of interest and salience information, our method exploits deep learning technique to provide comprehensive photo collage representation, which in turn can benefit the output of high-quality collages.

Aesthetic evaluation for image is extensively examined and successfully employed in multiple tasks like image quality assessment [12], [13], image cropping [14], [15] and image composition [16], [5], benefiting from the powerful feature representation of deep neural networks. Typical image aesthetic assessment approaches rely on multi-patch representation [6], [7], [17], [8], which represents each image with multiple cropped patches to learn global and local detail information simultaneously and are proven to be useful. Although many succeeding works [17], [8] 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, which can provide comprehensive photo collage representation.

Reinforcement learning methods were applied to multiple computer vision tasks in recent years, including image cropping [18], [19], image enhancement [20] image restoration [21] and object tracking [22]. Such works 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, the collage generation is decomposed into interpretable steps and modeled as reinforce-

ment learning process for the first time. Figure 1 illustrates 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. Since 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, aiming to explore general aesthetic attributes among images.

The network architecture is composed of nine layer convolution blocks, resembling the object detection architecture Single Shot MultiBox Detector[23]. Due to high expense of manual annotations and lack of high quality photo collage datasets, we pretrain our main aesthetic network on a large scale image aesthetic dataset (CPC dataset) 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 better adaptation in collage. On this basis, the attention layer is designed for structural aesthetic feature representation of a collage.

Since the most concerned part for a collage is the composition and color collocation of composed subimages, the fusion module shifts more focus to the composition among adjacent subimages instead of local parts from single image. Specifically, the fusion module set standards for feature selection extracted from a candidate patch with an area proportion greater than η in order to discover better composition quality and more harmonious content placement, which is calculated as follows:

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

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

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 which highlight the important images, we introduce the Rule of Central frequently used in photography and pay specific attention to patches close to the center of the collage for further aesthetic quality improvement. Since the patch information is extracted sharing the same weights, we inherently add the central rules in the multi-patch fusion process via attention mechanism, which is formulated as:

$$l_i = \left\| \frac{y_i}{h_c} - y_c \right\|_2 + \left\| \frac{x_i}{w_c} - x_c \right\|_2 \quad (2)$$

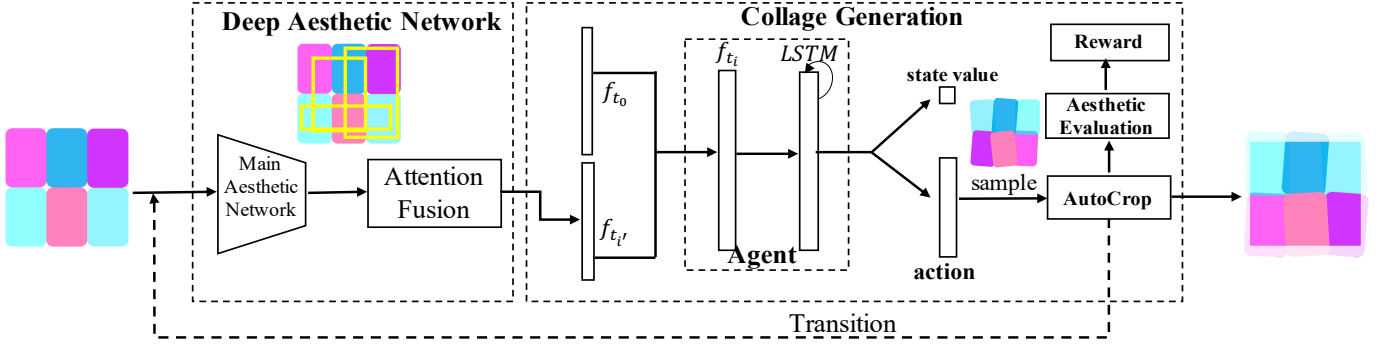


Fig. 1. The proposed network architecture for automatic collage generation. First, the aesthetic network takes in concatenated pictures best suited for the aspect ratio specified canvas as initialization. At each step, the deep aesthetic network first extracts both general aesthetic feature and structural 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 I. 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.

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

(y_i, x_i) denotes the center coordinate of i -th patch, (h_c, w_c) correspond to the canvas size, and (y_c, x_c) is chosen $(0.5, 0.5)$ to represent the focus center on canvas. The weighting factors is therefore represented as $\alpha(C) = [\alpha_1, \alpha_2, \dots, \alpha_n]$. Finally, combining the general and structural aesthetic feature, the representation of the holistic collage is computed as:

$$F(C) = \alpha(C) \cdot f(C) \quad (4)$$

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.

Figure 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, whereas the reward function is designed considering 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 and Action 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 $s_t = \{o_0, o_1, \dots, o_t\}$ includes all the past observations. To take advantage of the historical experience, a long short-term memory (LSTM) unit is added in the agent, assisting in making a better policy.

Inspired from the quick initialization in [4] which optimizes the collage from layout to details, we customize the action space for collage generation into two categories to adjust the global layout and local details with different attributes.

The global layout has substantial influences on the composition quality and is essential for a visually satisfying collage. Inspired from long distance image dragging in manual collages, our agent switches one image pair at each step before reaching the max step. The switch action could be interpreted as changing the order of input images. It affects the global layout through optimizations of image adjacency collocations and is done in multiple consecutive steps to adjust positions of images of more importance (like group photos) in a collage. Additionally, as 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 transforming process and output the current layout.

For detail adjustments, the agent operates on each individual image. Inspired from the state variables defined on the image set in [4], the detail adjustment attributes are designed with spatial position, placement order and orientation angles.

Due to the complex search space of spatial search space, it is clearly inefficient to directly operate the coordinates of each individual image. 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 within one episode. Subsequently, the agent adjusts the position of each image by reference to the neighboring image. The operations for layer are useful to display highlights and hide irrelevant area, while the rotation operation is designed to satisfy natural preference and improve the visual impression of the collage results.

2) AutoCrop Module:

The AutoCrop Module adapts 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. To be specific, 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 the view selec-

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.

TABLE I: The action space design considers global layout and local details. The global layout actions aim to improve the composition quality through image pair switching before the termination or max step. The local detail actions include relative spatial position, orientation angle and placement order (layer) for fine-grained collocation of sub-images.

tion is completed in the evaluation network and the cropped collage with the highest score is transitioned to the next step. As a result, the agent is encouraged to choose actions that progressively avoids losing salient information and meanwhile suppress the blank space on the canvas.

3) Reward Design:

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

Since the major challenge for analysing photo collage lies in the structural complexity, we assess the subjective quality with composition quality. Specifically, we resort to aesthetic evaluation network and propose using the aesthetic proposal number satisfying the aesthetic selection standard as representation for the subjective score of the collage C , which is denoted as $s_a(C)$. Intuitively, collage results with more aesthetic box number are indicators of higher aesthetic quality. For the objective score, since the 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:

$$s(C_t) = \lambda_a s_a(C_t) - \lambda_b s_b(C_t) \quad (5)$$

where the λ_a and λ_b stands for the weights for the aesthetic and the blank loss item, respectively. After each step, the difference of 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. In order to increase the aesthetic quality whilst suppress the blank space, the agent is granted a positive reward if the score increases and a negative reward otherwise.

$$r'_t(C_t) = s(C_{t+1}) - s(C_t) \quad (6)$$

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

$$r_t(C_t) = r'_t(C_t) - 0.01 * (t + 1) \quad (7)$$

4) 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 with 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 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:

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

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

where the optimization goal for 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 can encourage the agent to learn more flexible policies.

IV. EXPERIMENT

A. Experiment Settings

Training Dataset. We train our model on the Hollywood2 [24] movie dataset, which is composed of 12 classes of human actions and 10 classes of scenes. The dataset provides comprehensive realistic movie scenes with challenging settings in video format. We generate key frames for 44 videos chosen from the first 100 human action videos and resize them to 540 x 900. Each video consists of 40 randomly sampled sets of images with different numbers for composing the final collage. A total of ~2,000 image sets are included in the training set, from which our agent learns robust policies from videos with diverse quality.

Evaluation Dataset. To prove the scalabilities of the proposed model, we test it on 10 videos from Hollywood2 dataset and three movies from the MPII Movie Description dataset[25](LSMDC3), including "the queen", "up in the air", "pride and prejudice". The LSMDC dataset provides image collections extracted from sequential time clips. To make meaningful collages and meet the needs of real world scenarios, the evaluation image series in LSMDC3 are chosen from the same context scene and duplicated frames are removed within the same time period. A total of ~400 and 600 image sets are included in the above two test sets, respectively.

Implementation Details. The aesthetic network and the RL network are implemented with PyTorch [26] on Ubuntu 16.04. Our aesthetic evaluation is outputted by View Proposal Network pretrained on the CPC dataset [5], which is an aesthetic ranker with advanced image aesthetic evaluation accuracy. To stabilize the training process, the evaluation and deep aesthetic network share the same feature-extracting unit.

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 [27] optimizer. The learning rate and weight decay are set to $1e-3$ and $1e-5$. In the A2C algorithm, we set the discount factor to

0.99 and the entropy weight to 0.01. We construct the LSTM unit with four layers in the agent and set the weight of the aesthetic score and blank loss to 1 and 0.01 respectively.

Evaluation Metrics. To assess the quality of the collage results comprehensively, the aesthetic score is developed to describe the overall quality of the collage results, which computes as

$$F(C) = \sum_i^N s(P_i) \cdot f(P_i) \cdot \delta\left(\frac{s(P_i)}{s(C)} > \eta\right) \quad (10)$$

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 with more aesthetic box number are indicators of higher aesthetic quality. The 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 10, the high-quality collage accumulates more aesthetic score over local parts, by means of more balanced compositions, less occlusion along boundaries or fewer blending artifacts.

B. Quantitative Evaluation

To assess the effectiveness of the proposed model, we evaluate different methods quantitatively. We examine the effect of our action space design, which adjusts the global layout and local details of a collage in the first section, and compare the proposed model's aesthetic score with that of several competing methods in the second section.

Methods	Aesthetic Score
Instagram Layout [28]	79.83
Shape Collage [29]	88.22
Circle Packing Collage [1]	63.1
Picture Collage [4]	83.2
AutoCollage [3]	103.6
Ours (w/ AutoCrop)	104.8
Ours (w/o attention)	106.1
Ours	110.6

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

Method	Evaluation Input No.	Proposal Number			Aesthetic Score		
		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 II: 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 I.



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 saliency 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.

1) Evaluation of Aesthetic Quality:

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

The competing methods include (a) AutoCollage [3], which creates a collage of representative elements from a image set and develops a sequence of optimization steps for collage generation; (b) Circle Packing Collage [1], which partitions a canvas using the importance of regions of interest from input images; (c) Picture Collage [4], 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[28], which is an app developed by Instagram that combines multiple photos into one single image with predefined templates; and (e) Shape Collage[29], which is an automatic photo collage maker that allows to make shape or blending collages of family photos in a harmonious way with more flexible templates.

To generate results from the same test set in volume, 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. As Picture Collage also achieves competitive results, we reimplement its quick initialization process to make comparisons.

From Table III and IV, we can see our method achieves consistently higher aesthetic score results than competing methods on both video and image datasets.

Methods	Aesthetic Score
Circle Packing Collage [1]	108.84
Shape Collage [29]	114.68
AutoCollage [3]	132.65
Ours (w/ AutoCrop)	135.11
Ours	138.86

TABLE IV: Quantitative evaluation of different collage methods on the LSMDC3 dataset.

2) Evaluation of Action Space Design:

To evaluate the effectiveness of proposed action space design in Table I, we examine the effect of layout adjustment (C1) and local detail optimization (C2) on the improvement of collage aesthetic quality. As Picture Collage also generates collages with spatial coordinates, rotation angles, layer indices

and performs optimization through handcrafted energy terms, we reimplement quick initialization of picture collage as baseline method and quantitatively compare the quality of the intermediate layout and final results of both methods and evaluate them based on their proposal number and aesthetic score. The visual comparison of different action sets are illustrated in Figure 2.

For global layout adjustment, we use the layout initialization from picture collage as initialization to our network. We compare the layout results of our proposed network with those of baseline method for multiple different numbered inputs. As shown in Figure 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 details refinement, both methods are initialized with the same layout for fairness. Compared with the baseline method which pays more attention to salience constraints, our method receives feedback from subjective and objective factors. Table II shows that our results can achieve better improvement compared with salience-based optimizations.

C. Visual Comparisons with Existing Methods

We qualitatively compare our methods against several competitors with three typical collage styles(described in Section 4.2.1), i.e., blending style, overlay style and template style. The visual comparisons are displayed in Figure 4.

The blending style based methods include AutoCollage [4] and Circle Packing Collage [1]. As shown in Figure 4, AutoCollage shows multiple artifacts and loses salient information along the boundaries. Also, AutoCollage fails to highlight the visual focus when it comes to complex scenes. Circle Packing Collage loses most salient information and generates confusing boundaries in most cases. Picture Collage [4] introduces the



Fig. 3. Results of the proposed model on four mainstream aspect ratio specified canvases(i.e., "1:1", "16:9", "4:3", "3:4").



Fig. 4. Visual comparisons on the LSMDC3 dataset with competing methods. The aesthetic score using metric in Section 4.1 is labeled in the upright corner of every individual collage from different method. Artifacts along the boundaries of images are annotated with red boxes. In both scenarios, the AutoCollage presents multiple artifacts along the blending borderline, the Circle Packing Collage loses most salient information, and the Picture Collage loses salient information on the canvas border and struggles at highlighting the visual focus. 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.

overlay style to avoid blending artifacts, however, it is based on salience energy optimization and is unable to generate aesthetic photo collage with good composition quality. Instagram Layout[28] and Shape Collage [29] are template style based methods. The former method generates collage according to the input order and the latter randomly generates photo collage with rich grid templates. However, both methods relies on more user interaction in practice.

Compared to the competing models, our method exhibits the strengths of better composition quality whilst preserving the local details. The irrelevant area is greatly decreased with the occlusion while the salience object is well persevered. Plus, the results of our proposed method have clear boundaries and avoid artifacts.

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

D. Ablation Study

To prove the effectiveness of the proposed components, we design ablation experiments to prove the function of the attention fusion module and the AutoCrop module and verify the reasonableness of the evaluation metric.

1) Attention Fusion Module:

We examine the influence of attention fusion module on the aesthetic quality of generated collage results and research the effect of the attention mechanism.

As shown in Table III, after we remove the attention layer, the aesthetic score is observed to decrease by 4.5. Since the attention layer shifts more focus to the center of a collage, it is an implicit utilization of central rules and thus can enhance the subjective quality of collage results.

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. As 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 concerned area on the aspect ratio specified canvas.

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

Results in Table III, IV also show that the concerned content is considerably preserved without being cropped by canvas borders and demonstrate the effectiveness on both video and image datasets.

3) Evaluation Metric:

We explore the plausibility of the evaluation metric proposed in Formula 10. 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 view brings noisy signals and causes oscillation to the training process and the performance witnesses a reduction by 1.5% and 4.85%, respectively. Also, increasing the threshold to 70% reduces the structural information and leads to performance drop by 1.25%.

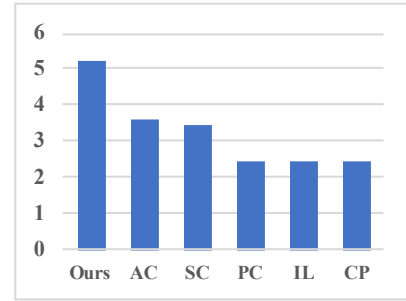


Fig. 5. User study on different photo collage methods. User study reports results in line with the aesthetic score in Table III, which is a further proof of the effectiveness of proposed evaluation method.

E. User Study

Apart from quantitative collage evaluations, to assess the collage methods subjectively, we conduct a user study using a questionnaire. With reference to [3], we prepare 24 groups of photo collages by six different methods and invite 20 users not involved in the work to rank the collages with different methods using the same input from top to worst (i.e., 6 to 1) to find visually pleasing collages. The collages are arranged in random order to avoid biases. Table 5 lists the average scores of each method. The proposed method received high evaluations, which prove its effectiveness. Moreover, the user study reports results in line with the aesthetic score in Table III, which is a further proof of the effectiveness of proposed evaluation method.

V. CONCLUSION

In this paper, a novel pipeline for automatic photo collage generation is proposed. Inspired by manual collages, collage generation is decomposed into interpretable steps and modeled as RL for the first time. The attention fusion module embedded in the deep aesthetic network is proposed to overcome the lack of training data and provide a comprehensive feature representation for the photo collage. Moreover, the AutoCrop module is proposed to inherently generate an aspect ratio specified collage, which makes the application scenario more flexible. Experiments on Hollywood2 and LSMDC3 video dataset demonstrate the superiority of the proposed model, and a user study further proves the effectiveness of the subjective evaluation and our method.

REFERENCES

- [1] 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.
- [2] 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.
- [3] C. Rother, L. Bordeaux, Y. Hamadi, and A. Blake, "Autocollage," *ACM transactions on graphics*, vol. 25, no. 3, pp. 847–852, 2006.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] C. B. Atkins, "Blocked recursive image composition," in *ACM international conference on Multimedia*, 2008, pp. 821–824.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] 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.
- [19] —, "Fast a3rl: Aesthetics-aware adversarial reinforcement learning for image cropping," *IEEE Transactions on Image Processing*, vol. 28, no. 10, pp. 5105–5120, 2019.
- [20] 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.
- [21] 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.
- [22] 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.
- [23] 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.
- [24] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 2929–2936.
- [25] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville, and B. Schiele, "Movie description," *International Journal of Computer Vision*, 2017. [Online]. Available: <http://resources.mpi-inf.mpg.de/publications/D1/2016/2310198.pdf>
- [26] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8026–8037.
- [27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [28] I. Layout. (2015) Instagram layout. [Online]. Available: <https://play.google.com/store/apps/details?id=com.instagram.layout>
- [29] V. Cheung. (2013) Shape collage. [Online]. Available: <http://www.shapecollage.com/>