Supplementary Materials for "SoftCollage: A Differentiable Probabilistic Tree Generator for Image Collage"

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

1. MATHEMATICAL DETAILS ON OUR APPROACH

1.1. Likelihood $L(\tau;\theta,\pi)$

We define $L(\tau;\theta,\pi)$ as the likelihood of a standard collage tree τ given canvas width w, canvas height h and image collection $\{I_i\}$, where π is the nearest neighbor policy (NNP) and θ is the parameters of the tree generator. $L(\tau;\theta,\pi)$ can be calculated using our probability space as

$$\begin{aligned} L(\tau;\theta,\pi) &= p(\tau|w,h,\{I_i\};\theta,\pi) \\ &= p(\tau|\tau_\theta,w,h,\{I_i\};\theta,\pi)p(\tau_\theta|w,h,\{I_i\}) \\ &= p(\tau|\tau_\theta,w,h,\{I_i\};\theta,\pi) \\ &= p(\tau|\tau_\theta) \end{aligned} \tag{1}$$

where τ_θ is the probabilistic collage tree ,*i.e*. PCtree.

1.2. Gradient $\nabla_\theta\overline{F_\theta}(\tau;\pi)$

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

$$\begin{aligned} \nabla_\theta\overline{F_\theta}(\tau;\pi) &= \nabla_\theta\Big(\sum_{\tau\sim L(\tau;\theta,\pi)} F\big(g(\tau)\big)L(\tau;\theta,\pi)\Big) \\ &= \sum_{\tau\sim L(\tau;\theta,\pi)} F\big(g(\tau)\big)\nabla_\theta L(\tau;\theta,\pi) \\ &= \sum_{\tau\sim L(\tau;\theta,\pi)} F\big(g(\tau)\big)L(\tau;\theta,\pi)\nabla_\theta \log L(\tau;\theta,\pi) \\ &\approx \frac{1}{M}\sum_{i=1}^{M} F\big(g(\tau_i)\big)\nabla_\theta \log L(\tau;\theta,\pi) \\ &= \nabla_\theta\Big(\frac{1}{M}\sum_{i=1}^{M} F\big(g(\tau_i)\big)\log L(\tau;\theta,\pi)\Big) \\ &= \nabla_\theta\Big(\frac{1}{M}\sum_{i=1}^{M} F\big(g(\tau_i)\big)\log p(\tau|\tau_\theta)\Big) \end{aligned} \tag{2}$$

where M is the number of sample τ_i.

1.3. Inference

At inference stage, the optimal collage tree τ^* is determined by the maximum likelihood method as

$$\tau^* = \arg\max_{\tau} p(\tau|\tau_\theta) \tag{3}$$

Notably the structural parameter inference of each node is independent of each other. Moreover, the nodes in the PCtree and standard collage tree are in one-to-one correspondence. Thus, given a node $\tilde{n}$ in the PCtree, the corresponding node n in the standard collage tree is determined once the cut type c_n is predicted via

$$c_n^* = \arg\max_{c_n\in\{\text{"H"},\text{"V"}\}} p_n^{(\mathbb{1}\{c_n=\text{"V"}\})}\big(\tilde{n}\big) \tag{4}$$

After the node inference, given a node $\tilde{n}$ with sub-nodes $\tilde{n_i}$ and $\tilde{n_j}$ in the PCtree, the corresponding nodes n, n_i and n_j in the standard collage tree are determined. Hence, the

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

edge connection relationship is determined by

$$l_n^* = \underset{l_n \in \{n_i, n_j\}}{\arg\max}\ p_e^{(0)}\left(\tilde{l_n}, \{\tilde{n_i}, \tilde{n_j}\} \setminus l_n\right) \tag{5}$$

$$r_n^* = \{n_i, n_j\} \setminus l_n^* \tag{6}$$

where l_n is the left node of node n, r_n is the right node of node n, $\tilde{l_n}$ is the PCtree node corresponding to l_n and $\{n_i, n_j\} \setminus x$ denotes the element that is not equal to x in the binary set $\{n_i, n_j\}$.

2. EXPERIMENTAL DETAILS AND MORE RESULTS

2.1. The Image Collection Sampling Framework

The image collection sampling framework is designed to generate AIC from the ICSS. This framework samples train set and test set of AIC respectively from train set and test set of ICSS. Each image collection sampled by the framework satisfies the following constraints:

1. Images in the collection share the identical theme of the ICSS.
2. The collection includes images from at least two categories.
3. There are at least two images in each category in the collection.
4. The category distribution of the collection conforms to uniform distribution and is not biased by the prior category distribution in the ICSS.

where the second and the third constraints are established for acquiring effective M_n value.

Our framework samples image collections under each collection size and theme for multiple times through the enumeration method. Hence, our framework has two parameters, where one is the list of the collection sizes defined as $\{s_i\}$ and the other is the maximum number of sampling defined as T. The framework is described in Algo. 1, where $\lceil \binom{s}{\mid C \mid}/2 \rceil$ in the seventh line is introduced to prevent sampling identical collections when s is close to or equivalent to $\mid C \mid$ and the constraints in the tenth line is used to satisfy the second and the third constraint above. To solve the problem of Eq. (7) under the above four constraints, we develop a distribution-balanced sampling algorithm, which is parameterized by s, *i.e.* collection size, and $\{c_i\}_{m_{th}}$, *i.e.* list of the number of images of each category, where m_{th} denotes the number of categories in theme th. The algorithm is described in Algo. 2, where notably in the second line $\min(m_{th}, \lfloor s/2 \rfloor)$ is used to satisfy the third constraint and $\max(2, \lceil \frac{s}{\min_{1 \le i \le m_{th}} c_i} \rceil)$ is used to satisfy the second constraint and to ensure that there are always at least s images in any m categories. Notably n should not be less than 4 for using this algorithm.

We set $\{s_i\}$ to $\{10, 15, 20, 25, 30, 50, 100\}$, and T is set to 10 and 2 for the train set and the test set respectively. As a result, the train set has 562 image collections including 18535 images and the test set has 62 image collections including 1260 images.

2.2. Implementation Details

We implement the proposed framework using the PyTorch toolbox [8] on one GeForce RTX 3090 GPU. We set the canvas width w to 900 and height h to 600. The ResNet-50 [4] pre-trained on the ImageNet [3] with d_{bb} equaling 2048 is adopted as the backbone network in our feature extractor. The information embedding hyperparameters d_w and d_h are both set to 4, and d_{ar} and d_{inf} are both set to 128. We set d_Q and d_K both to 1024, and set $d_V = d_{bb} + d_{inf}$. We set d_1 and d_2 to 512 and 2 respectively according to [6]. With respect to the loss function, we set M to 1000 and R_0 to 2. Moreover, r_1, r_2 and r_3 are set to 10^{-3}, 0.1 and 0.5 respectively. Notably we set s_p to 20.0 for F_p. The three criteria are added into F with $\lambda_r = \lambda_p = 1.0$ and $\lambda_a = 0.05$. To train our model, we use the Adam optimizer [5] with an initial learning rate α of 10^{-4} for each image collection and T_m is set to 100.

2.3. More Ablation Analysis

The self-attentive embedding in the fusion module. To study the necessity of the Eq (10) and Eq (11) in the main paper, we conduct some experiments of the ablated versions of the Eq (10) (*i.e.* average pooling and one fully connected layer) and the ablated version of the Eq (11) (*i.e.* identity transform) on the train set of the AIC. The results are presented in Tab. 1.

The distance criterion of the NNP. Since the distance criterion of the NNP has high effects on the model performance, we compare different distance criteria on the train set of the AIC, including Euclidean distance, standardized Euclidean distance and the cosine distance. The results are shown in Tab. 2. Notably we define the cosine distance of the two features f_i and f_j as $D_{cos} = 1 - \frac{<f_i, f_j>}{\|f_i\| \|f_j\|}$.

2.4. Human Evaluation

Firstly we carried out the 5-scale evaluation. Fifteen human raters participated in the evaluation and each of them was shown with 16 groups of collages. Each group includes four collages, *i.e.* one generated by our method and three by the baseline methods. The raters were asked to watch the collages for at least 20 s and rated them from 'Excellent' (4) to 'Bad' (0). To measure the gain in our method over the baselines, we also conducted the side-by-side evaluation. This comparative task is easier than 5-scale rating task for human and thus can produce more reliable results. Thirty raters participated in the evaluation and they were equally divided into three groups. Each group compares our

Version of the Eq (10)	Version of the Eq (11)	M_r	M_n
Original	Original	**1.086**	0.284
$A = softmax(W_s f'_{(i,j)})$	Original	1.142	0.280
$A = softmax\big(avg\_pooling(f'_{(i,j)})\big)$	Original ($d_2 = 1$)	1.282	0.273
$A = softmax\big(avg\_pooling(f'_{(i,j)})\big)$	Identity transform	1.657	**0.259**

Table 1. Ablation analysis of the self-attentive embedding in the fusion module on the train set of AIC. Here, $W_s \in \mathbb{R}^{d_2 \times d_V}$ is a learnable parameter.

Algorithm 1: The image collection sampling framework

Input: ICSS, $\{s_i\}, T$
Output: AIC

AIC $\leftarrow \{\}$
for *every theme* th *in the ICSS* **do** // enumerate theme
  Calculate m_{th}, *i.e.* the number of categories in theme th
  $C \leftarrow \{c_i\}_{m_{th}}$ // c_i is the number of images in the i-th category
  $\mid C \mid \leftarrow \sum_{i=1}^{m_{th}} c_i$
  for *every* $s \in \{s_i\} \cap [0, \mid C \mid]$ **do** // enumerate collection size
    Sample $t \sim DiscreteU\Big(1, \ldots, \min\big(\lceil \binom{s}{\mid C \mid}/2 \rceil, T\big)\Big)$ // *DiscreteU* is discrete uniform distribution and $\binom{x}{y}$ denotes combinatorial number
    $t_0 \leftarrow 0$
    repeat // sample for t times
      Randomly sample out $\{x_i\}_{m_{th}}$ to generate the image collection $S_{t_0}^{(th,s)}$ $s.t.$

$$\sum_{i=1}^{m_{th}} x_i = s,$$
$$x_i \in [0, c_i] \cap \mathbb{Z} \setminus \{1\},$$
$$\exists i \neq j \quad x_i x_j > 0 \tag{7}$$

      Add $S_{t_0}^{(th,s)}$ into AIC
      $t_0 \leftarrow t_0 + 1$
    until $t_0 = t$
  end
end
return AIC

Distance criterion	M_r	M_n
Euclidean	**1.086**	0.284
Standardized Euclidean	18.459	**0.213**
Cosine	1.179	0.291

Table 2. Comparison of different distance criteria of the NNP on the train set of AIC.

method with one of the three baselines. We showed every participant 16 pairs of collages. Each pair includes one collage generated by our method and one by the corresponding baseline. The raters were also required to observe for at least 20 s before giving the judge. Additionally, Fleiss' Kappa score is used to gauge the reliability of the agreement between evaluators.

2.5. Results

Due to the limited space of the main papar, the collages in the main paper are reduced to a small size and only some of the results are presented. To illustrate the superiority of our method more clearly, we here show more results in Figs. 1 to 11.

Algorithm 2: The distribution-balanced sampling algorithm

Input: $s, \{c_i\}_{m_{th}}$
Output: $\{x_i\}_{m_{th}}$
$\{x_i\}_{m_{th}} \leftarrow \{0\}_{m_{th}}$
Sample $m \sim DiscreteU\Big(\min\big(\max(2, \lceil \frac{s}{\min\limits_{1 \le i \le m_{th}} c_i} \rceil), m_{th}, \lfloor s/2 \rfloor\big), \ldots, \min(m_{th}, \lfloor s/2 \rfloor)\Big)$ // m is the number of the sampled categories and $DiscreteU$ is discrete uniform distribution
Randomly sample out $\{i_s\}_m$ from $\{1, \ldots, m_{th}\}$ // $\{i_s\}_m$ denotes the list of the sampled categories
for *every* $i_s \in \{i_s\}_m$ **do**
    $x_{i_s} \leftarrow 2$ // ensure that there are at least two images in each category
end
$k \leftarrow 2m$ // k denotes the number of the sampled images
repeat
    $\delta \leftarrow \{\}$ // δ denotes the list of the categories that have images for sampling
    for *every* $i_s \in \{i_s\}_m$ **do**
        if $x_{i_s} < c_{i_s}$ **then** // c_{i_s} is the number of images in the i_s-th category
            Add i_s into δ
        end
    end
    Randomly sample i_s^* from δ
    $x_{i_s^*} \leftarrow x_{i_s^*} + 1$
    $k \leftarrow k + 1$
until $k = s$
return $\{x_i\}_{m_{th}}$

References

[1] Collageit. online, 2019. https://www.collageitfree.com/. 5, 6, 7, 8, 9, 10

[2] V. Cheung. Shape collage. online, 2013. http://www.shapecollage.com/. 5, 6, 7, 8, 9, 10

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[5] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 2

[6] Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 2

[7] Xingjia Pan, Fan Tang, Weiming Dong, Chongyang Ma, Yiping Meng, Feiyue Huang, Tong-Yee Lee, and Changsheng Xu. Content-based visual summarization for image collections. *IEEE transactions on visualization and computer graphics*, 2019. 5, 6, 7, 8, 9, 10

[8] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019. 2

(a) SHP [2]

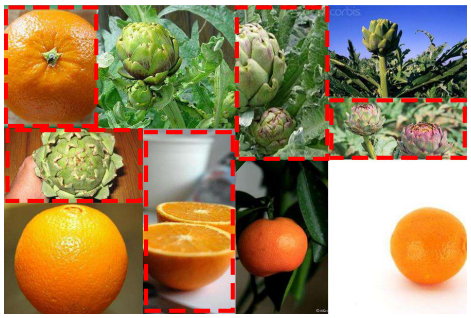
(b) CLT [1]

(c) VSM [7]

(d) Ours

Figure 1. Comparison of the collage results generated by different methods on the AIC. The image content occlusion and severe aspect ratio distortion are both highlighted by red dotted rectangle.

(a) SHP [2]



(b) CLT [1]

(c) VSM [7]

(d) Ours

Figure 2. Comparison of the collage results generated by different methods on the AIC. The image content occlusion and severe aspect ratio distortion are both highlighted by red dotted rectangle.

(a) SHP [2]

(b) CLT [1]

(c) VSM [7]

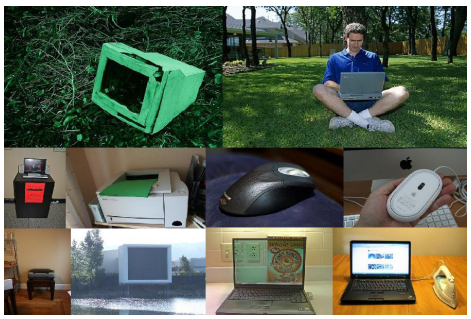
(d) Ours

Figure 3. Comparison of the collage results generated by different methods on the AIC. The image content occlusion and severe aspect ratio distortion are both highlighted by red dotted rectangle.

(a) SHP [2]

(b) CLT [1]

(c) VSM [7]



(d) Ours

Figure 4. Comparison of the collage results generated by different methods on the AIC. The image content occlusion and severe aspect ratio distortion are both highlighted by red dotted rectangle.

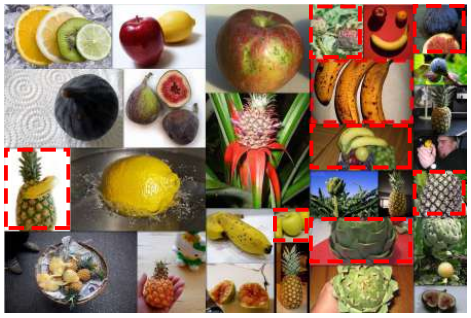
(a) SHP [2]

(b) CLT [1]

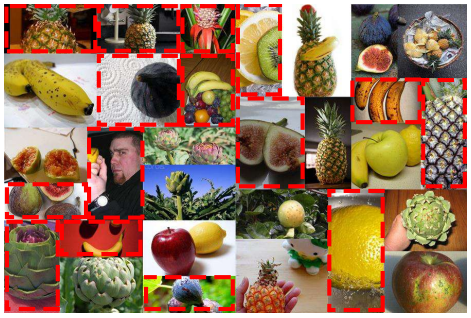
(c) VSM [7]

(d) Ours

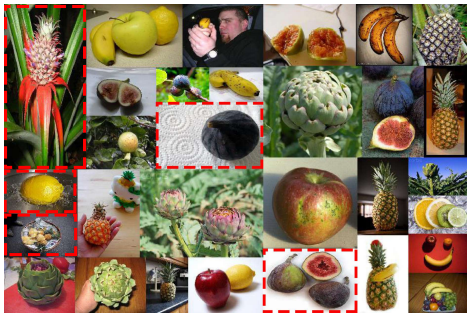
Figure 5. Comparison of the collage results generated by different methods on the AIC. The image content occlusion and severe aspect ratio distortion are both highlighted by red dotted rectangle.



(a) SHP [2]



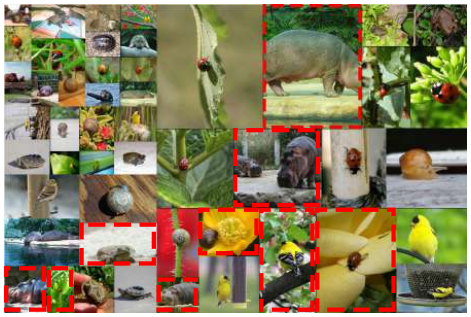
(b) CLT [1]



(c) VSM [7]

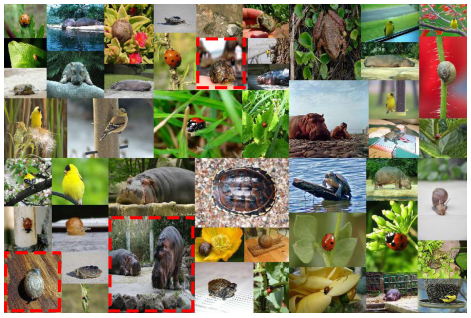
(d) Ours

Figure 6. Comparison of the collage results generated by different methods on the AIC. The image content occlusion and severe aspect ratio distortion are both highlighted by red dotted rectangle.

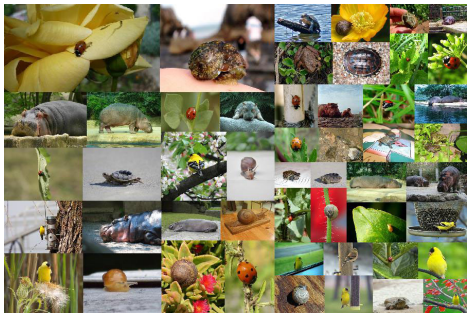


(a) SHP [2]

(b) CLT [1]

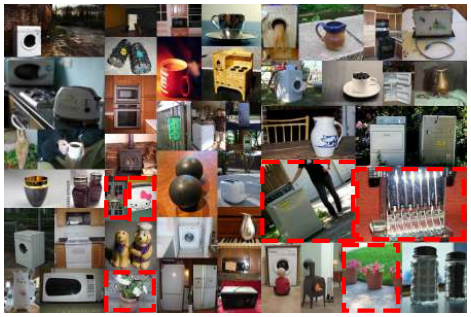


(c) VSM [7]



(d) Ours

Figure 7. Comparison of the collage results generated by different methods on the AIC. The image content occlusion and severe aspect ratio distortion are both highlighted by red dotted rectangle.



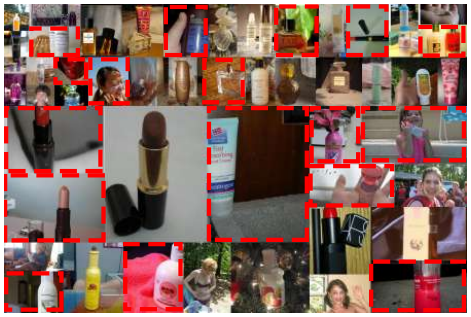
(a) SHP [2]

(b) CLT [1]

(c) VSM [7]

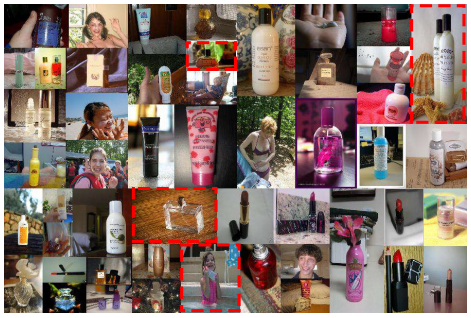
(d) Ours

Figure 8. Comparison of the collage results generated by different methods on the AIC. The image content occlusion and severe aspect ratio distortion are both highlighted by red dotted rectangle.

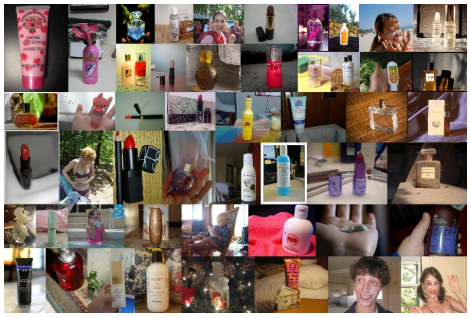


(a) SHP [2]

(b) CLT [1]



(c) VSM [7]



(d) Ours

Figure 9. Comparison of the collage results generated by different methods on the AIC. The image content occlusion and severe aspect ratio distortion are both highlighted by red dotted rectangle.

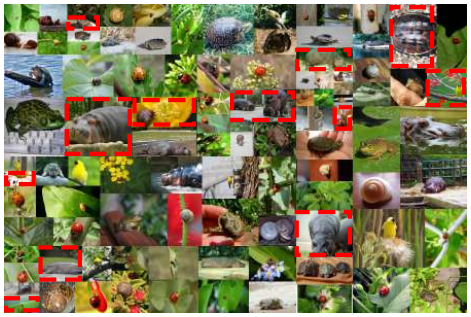
(a) SHP [2]

(b) CLT [1]

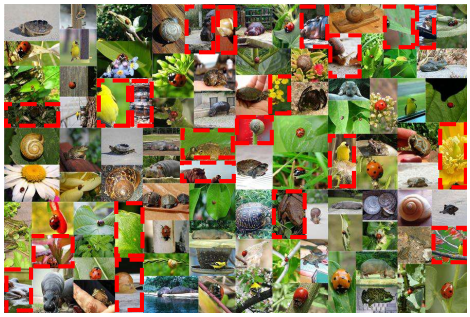
(c) VSM [7]

(d) Ours

Figure 10. Comparison of the collage results generated by different methods on the AIC. The image content occlusion and severe aspect ratio distortion are both highlighted by red dotted rectangle.



(a) SHP [2]



(b) CLT [1]

(c) VSM [7]

(d) Ours

Figure 11. Comparison of the collage results generated by different methods on the AIC. The image content occlusion and severe aspect ratio distortion are both highlighted by red dotted rectangle.