Neural Collage Transfer: Artistic Reconstruction via Material Manipulation

Ganghun Lee, Minji Kim, Yunsu Lee, Minsu Lee, Byoung-Tak Zhang
Seoul National University
{khlee, mjkim, yslee, mslee, btzhang}@bi.snu.ac.kr

In this supplementary material, we present additional results that can aid in a more detailed investigation of our method. Moreover, we quantitatively evaluate the availability of CLIP [41] on style and content retention measures. Proofs and verification of our proposed methods are also included.

A. Additional Results

A.1. Customized Results

Lots of hyper-parameters and possible environmental setups (*e.g.*, complexity thresholding, material images, scale adjustment, training strategy, action design, material ratio, image size) of the proposed method allows highly customized collage generation. We present several fine-tuned generated examples in Fig. 1, Fig. 2, and Fig. 3.

A.2. Comparision with Naïve Approach

To enhance the impact of our methods, we specially built a naïve collage generation algorithm. It divides the canvas area into several pieces of random Voronoi diagram, then searches materials for each area lowering MSE. The example results are illustrated in Fig. 4 and Fig. 5. They show why the naïve approach is likely to fail even though the pieces are densely divided. Choosing the right shape for the collage is essential to make fine quality, but the shapes used in the naïve approach were randomized. Determining the shape is too complicated to define a simple heuristic. Moreover, in Fig. 5, we can see that the searching method is very time-consuming.

A.3. Scale and Target Image Variation

We provide additional results for other target images, as illustrated in Fig. 6. The images were obtained in the process of the coarse-to-fine multi-scale sequence $\mathcal{U} = \{512, 256, 128, 64, 32\}$, and each number on the figure is the scale u in $\mathcal{U}$ where the intermediate result was obtained. The materials used in this figure were from Times magazine.

A.4. Complexity Sensitivity Variation

To enable a visual comparison of the effects of varying the complexity sensitivity parameter τ, we present additional collage results. In each case of $\tau = 1$ (Fig. 6), $\tau = 2$ (Fig. 7), and $\tau = 4$ (Fig. 8), areas with low complexity display more abstracted style when τ is high (please see the background of the boat.) The materials used in these figures are from Times magazine.

A.5. Material Variation

In order to demonstrate the effect of selecting different materials, we present additional neural collage transfer results using Newspaper [50] as an alternative material, as shown in Fig. 9. Newspaper contains a relatively faded color than Times, which also made the color tones of resulting collage images slightly faded.

A.6. Time Indicator Variation

To investigate the effect of t_m, namely time indicator, on multi-scale collage, we varied t_m as 0, 5, and 9 to produce collages using the same target image for the same number of timesteps as illustrated in Fig. 10. Since the agent was trained to make collage from the white canvas, it tended to make big scraps when t_m is low and small scraps when t_m is high. Please remember that we fixed $t_m = 9$ for a multi-scale collage in the main paper to make more detailed collages.

B. CLIP Score Verification

CLIP score was used for comparison of our method with NST. We state CLIP score is a proper measure for artistic style and content consistency by providing additional verification.

B.1. Style Verification

We first conducted the *style test* to verify CLIP's general style knowledge. We collected five images for each style category (photographic work, oil painting, collage, animation, sketch) and investigated the matching probabilities between the images and category texts. As in Fig. 11, CLIP

achieved 92% accuracy (71.2% of mean correct confidence) on this test and successfully distinguished each art style, proving its proper style recognition capacity. The failed case in row 3, column 1 predicted collage art as oil painting, but it is likely due to the image's mixed features of collage and abstract painting. Similar understandable confusions are seen in row 4, column 5, and row 5, column 4.

B.2. Content Verification

Next, we conducted the *content test* to verify CLIP's general content knowledge. We collected five images for each content category (airplane, banana, candy, dog, flower) and investigated the matching probabilities between the images and category texts. As in Fig. 12, CLIP achieved almost perfect accuracy (100%) on this test and successfully distinguished each content, proving its proper content recognition capacity.

C. Proofs

In this section, we provide proofs for equation (6) and (7).

C.1. Model-Based Soft Policy Evaluation

To derive *model-based soft policy evaluation* (6) from original *soft policy evaluation* (4), we can use the interchangeable relations (2) and (5). Considering (5) at time step $t+1$, following equation is also true:

$$Q(s_{t+1}, a_{t+1}) = r(s_{t+1}, a_{t+1}) + \gamma \mathbb{E}_{s_{t+2} \sim \mathcal{P}} \left[V(s_{t+2}) \right] \tag{13}$$

Substituting (2) into (4), we can obtain

$$\begin{aligned} J_Q = \mathbb{E}_{(s_t, a_t) \sim \mathcal{D}} [\tfrac{1}{2} (Q(s_t, a_t) - (r(s_t, a_t) \\ + \gamma \mathbb{E}_{s_{t+1} \sim \mathcal{P}} [\mathbb{E}_{a_{t+1} \sim \pi} [Q(s_{t+1}, a_{t+1}) - \\ \alpha \log \pi(a_{t+1} | s_{t+1})]]))^2]. \end{aligned} \tag{14}$$

By substituting (5) and (13) into (14), we can derive the objective function J_V (6) with respect to V instead of Q.

C.2. Model-Based Soft Policy Improvement

To derive the *model-based soft policy improvement* (7) from the original *soft policy improvement* (4), we can substitute (5) into (4) producing the objective function J_π (7) constituting of V instead of Q.

D. Differentiability Verification

For the sake of differentiability of the cut-and-paste function δ, two conditions should be met: (1) δ should not let the gradient become zero (*e.g.*, rounding), and (2) the action a_t can be approximately evaluated when C_t, M_t, C_{t+1} are given. Highly entangled action could violate condition (2), confusing the agent about the proper action usage. We tested the following sequence to verify if δ meets the above conditions using verification network $\mathcal{E}(\cdot; \phi)$.

1. Prepare example canvas C, material M, a constant input vector b, action a and corresponding next canvas $C' = \delta(C, M, a)$.
2. Input b into $\mathcal{E}$ to get output $\hat{a} = \mathcal{E}(b)$.
3. Get $\hat{C}' = \delta(C, M, \hat{a})$.
4. Update $\mathcal{E}$ with gradient descent $\phi \leftarrow \phi - \eta \nabla l_2(\hat{C}', C')$ (η is learning rate).
5. Repeat 2-4 to check if $\hat{a} \rightarrow a$.

If $\hat{a}$ does not change, δ may violate (1). If $\hat{a}$ does not converge to a, δ may violate (2).

Figure 1: *The Man of City*.

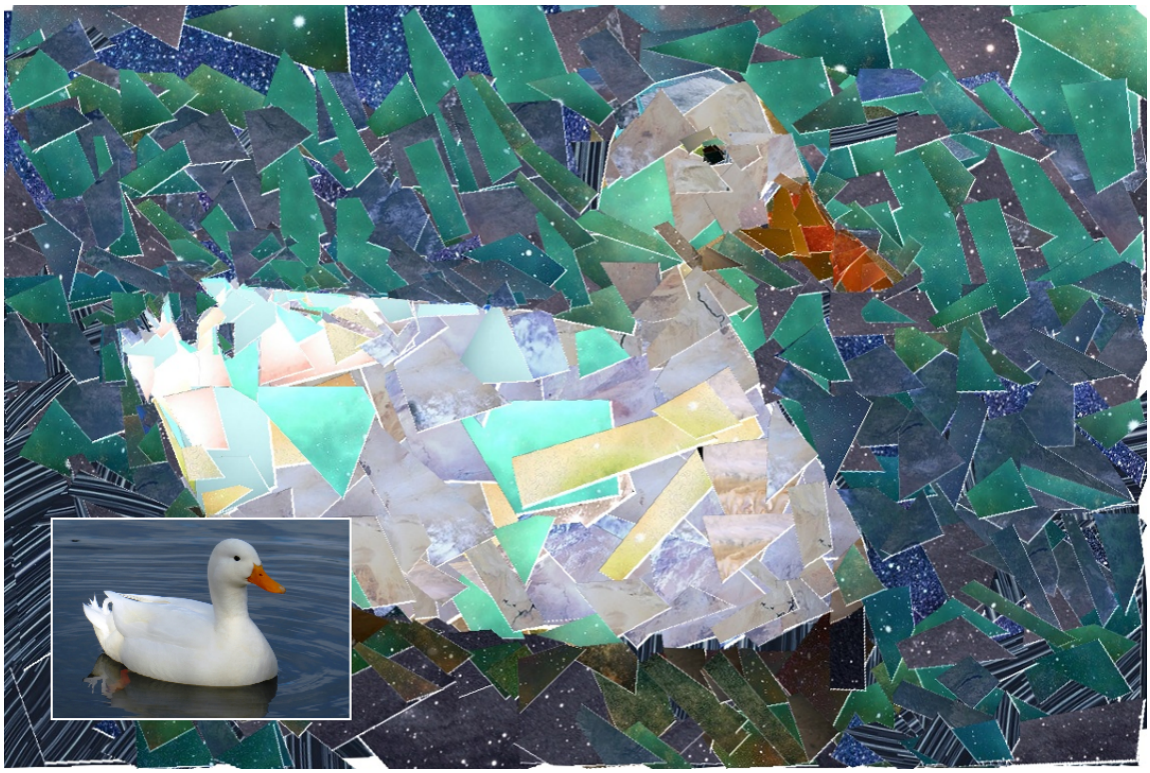


Figure 2: *Space Duck*.

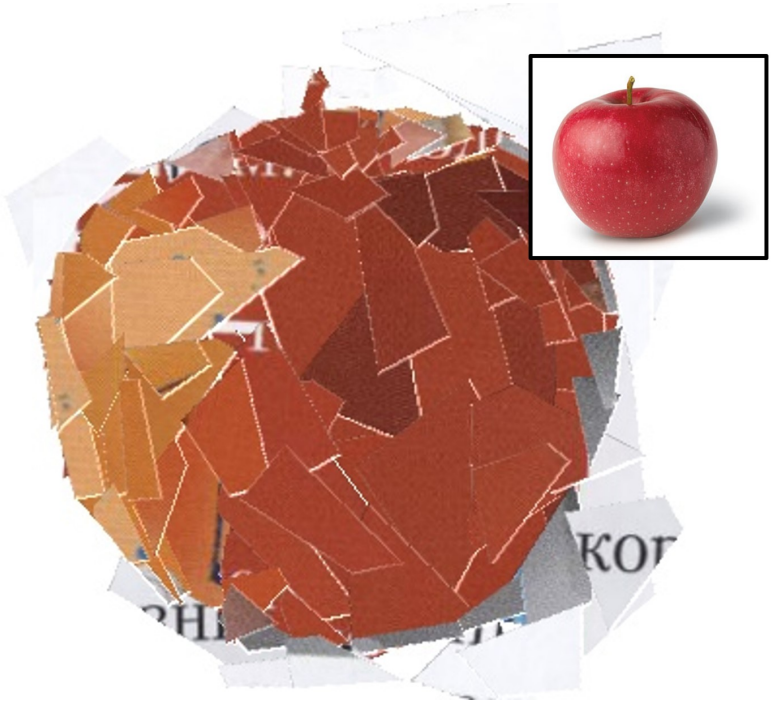


Figure 3: *Paper Apple*.

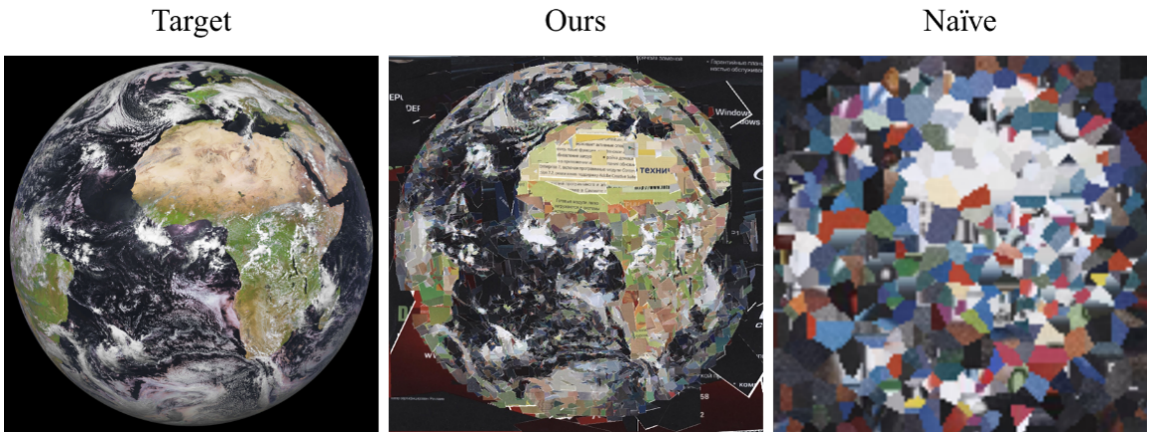


Figure 4: Earth, comparison with naïve approach

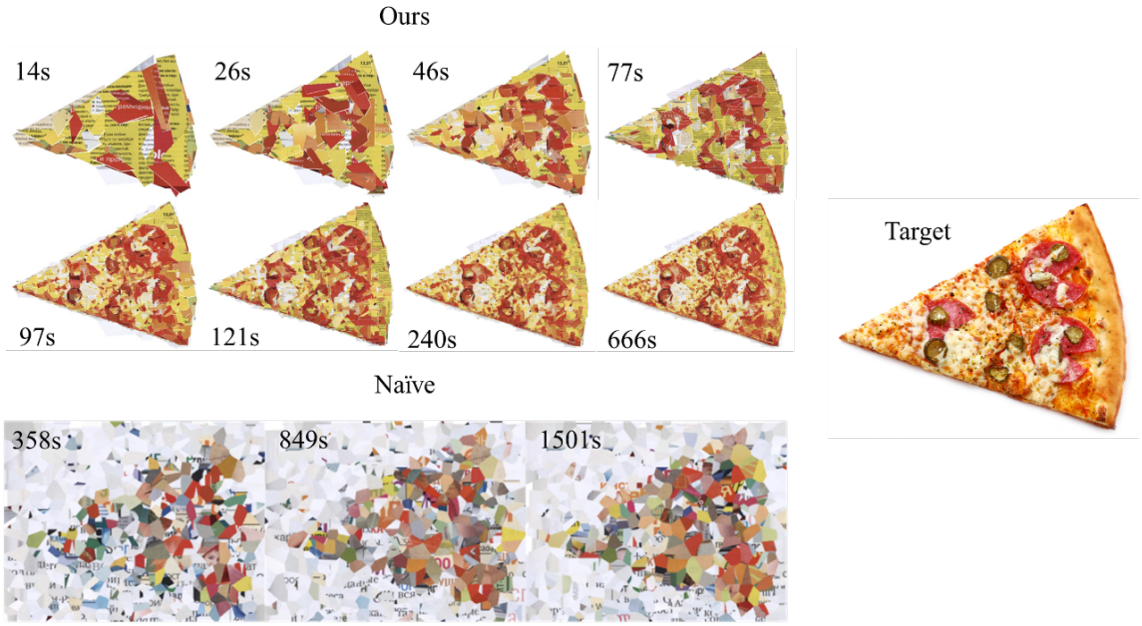


Figure 5: Pizza, comparison with naïve approach

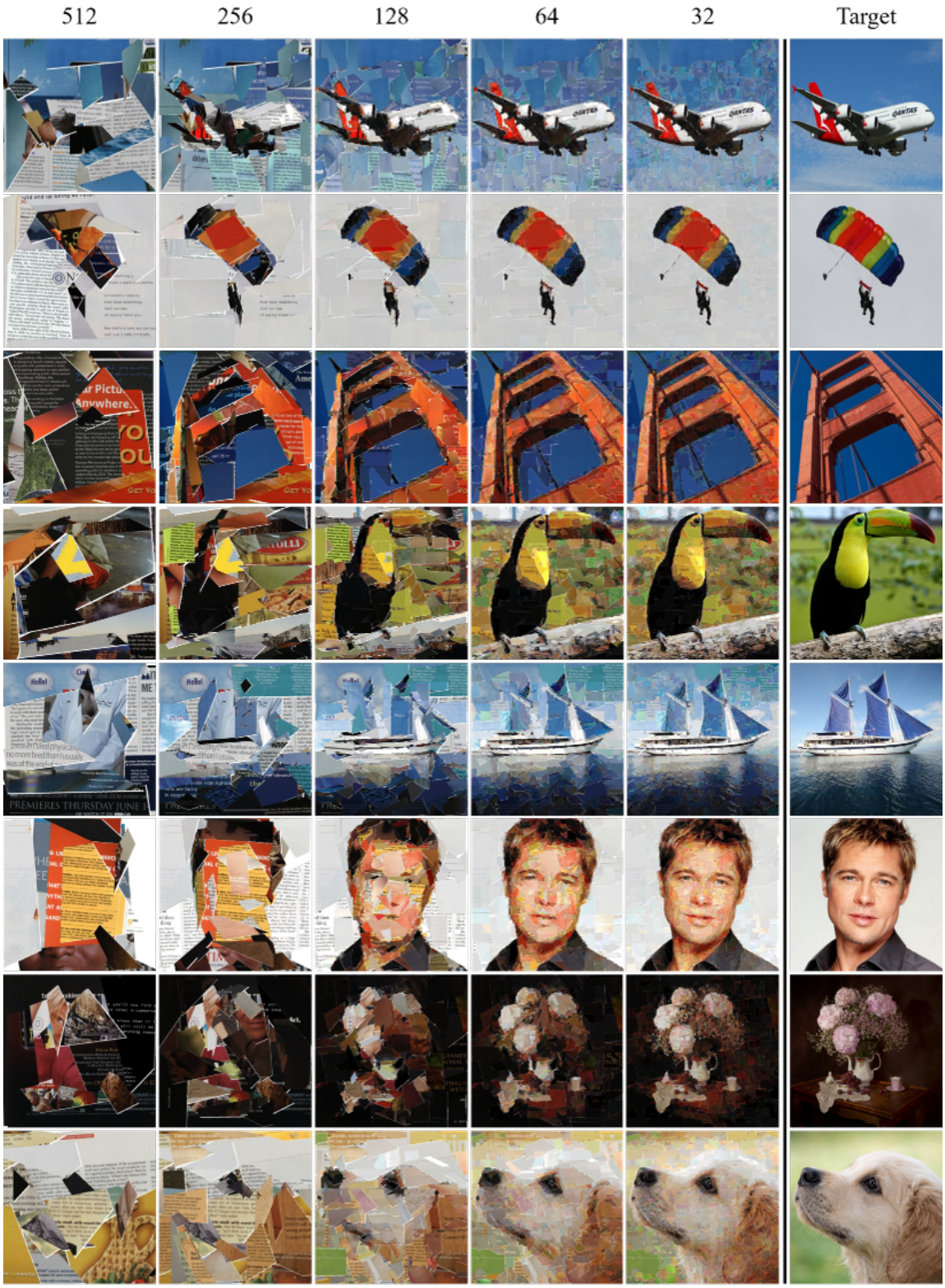


Figure 6: Collage generation sequences of ours ($\tau = 1$, materials: Times)

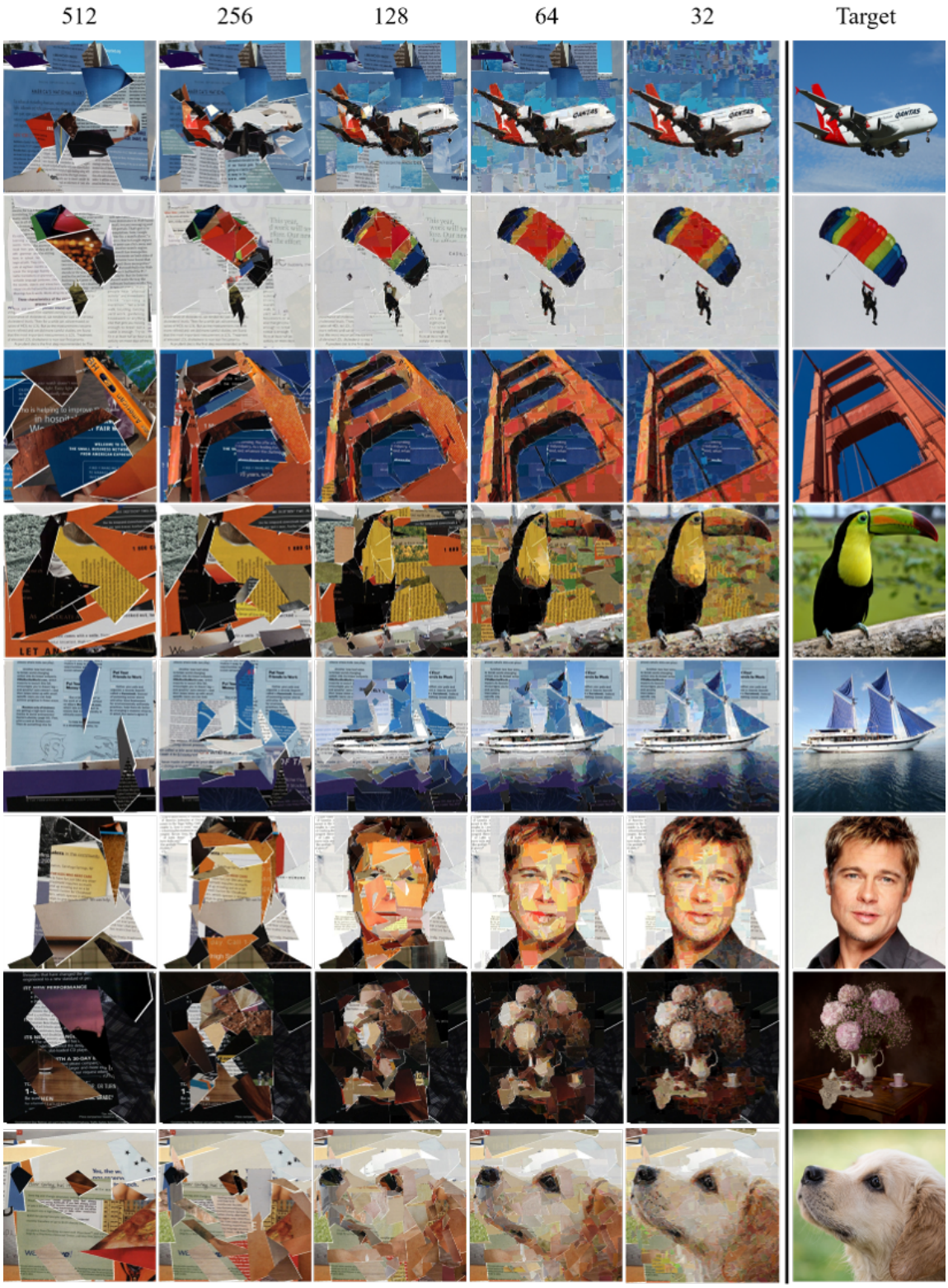


Figure 7: Collage generation sequences of ours ($\tau = 2$, materials: Times)

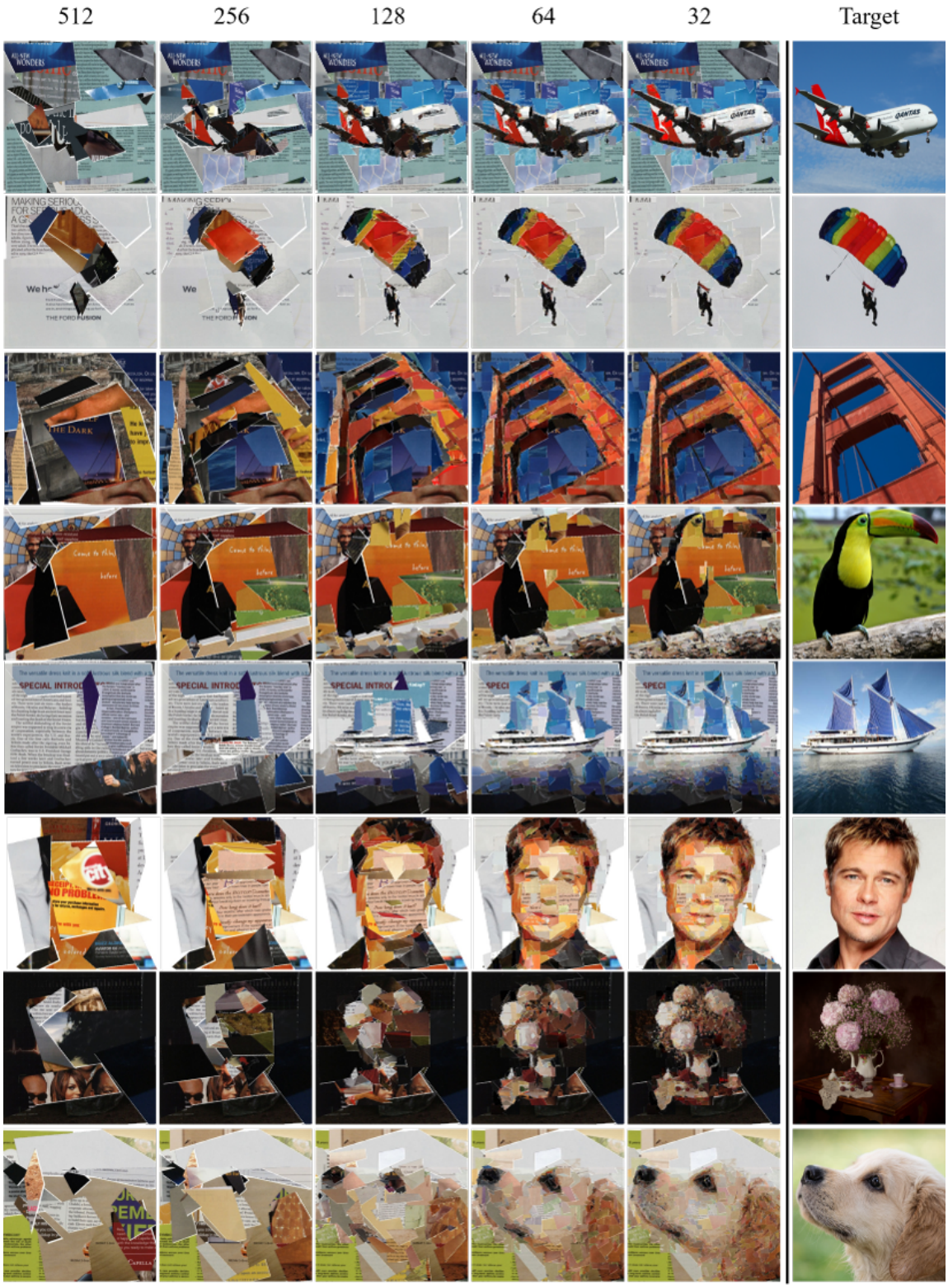


Figure 8: Collage generation sequences of ours ($\tau = 4$, materials: Times)

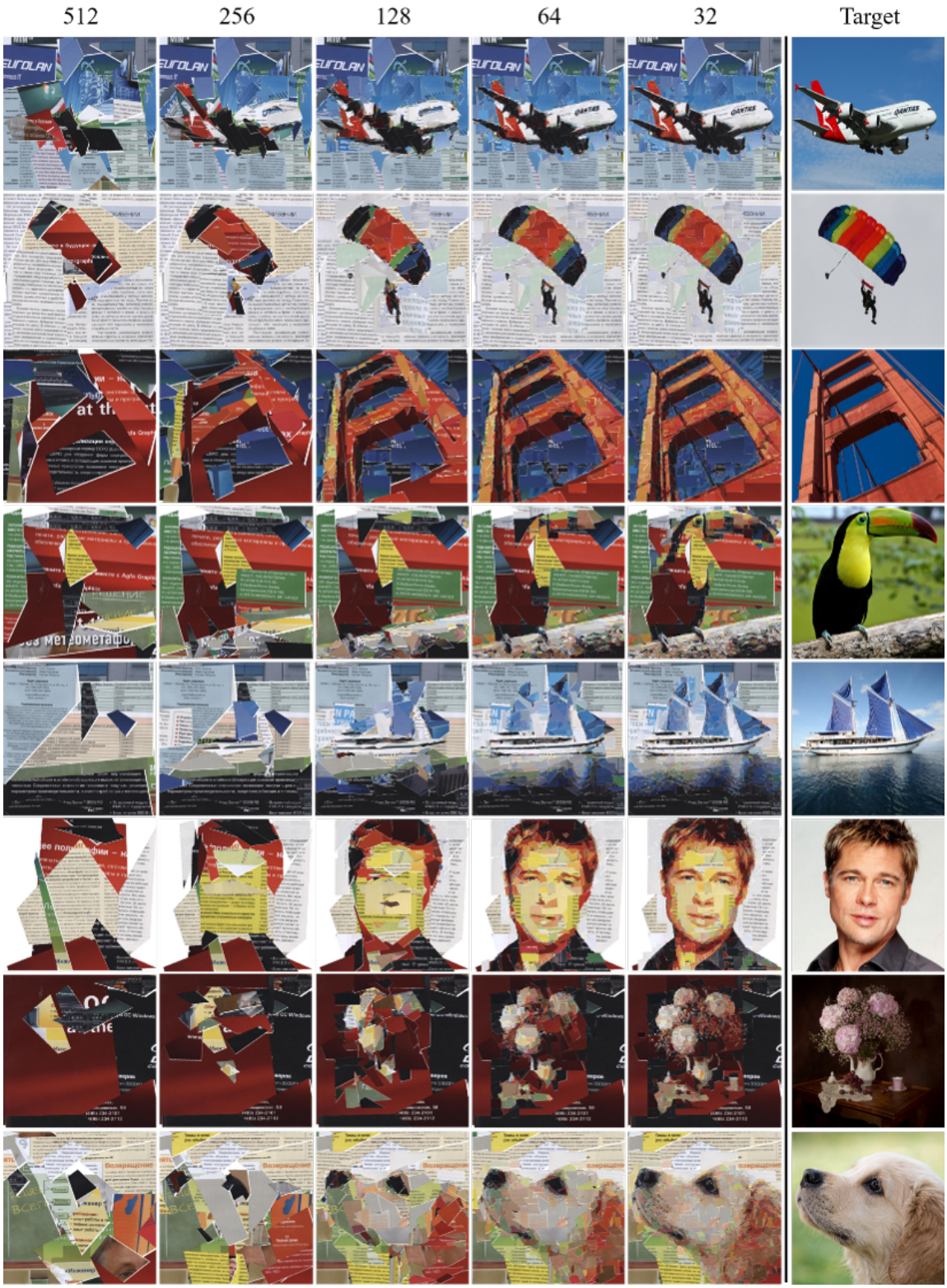
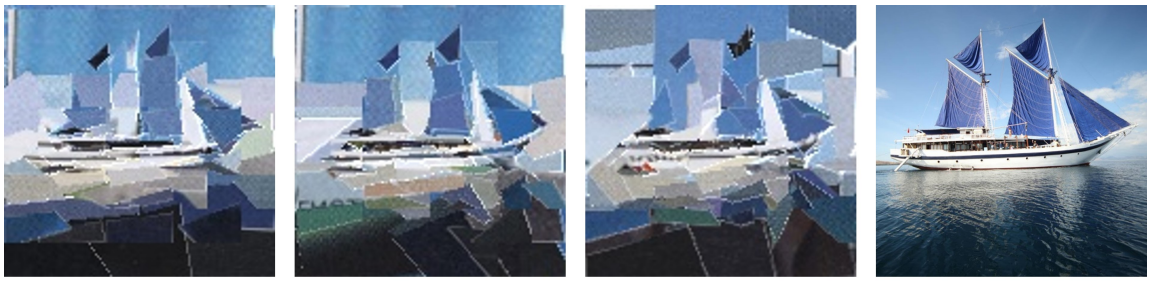


Figure 9: Collage generation sequences of ours ($\tau = 4$, materials: Newspaper)



$t_m = 0$ $t_m = 5$ $t_m = 9$ *Target*

Figure 10: Effect of varying t_m on multi-scale collage

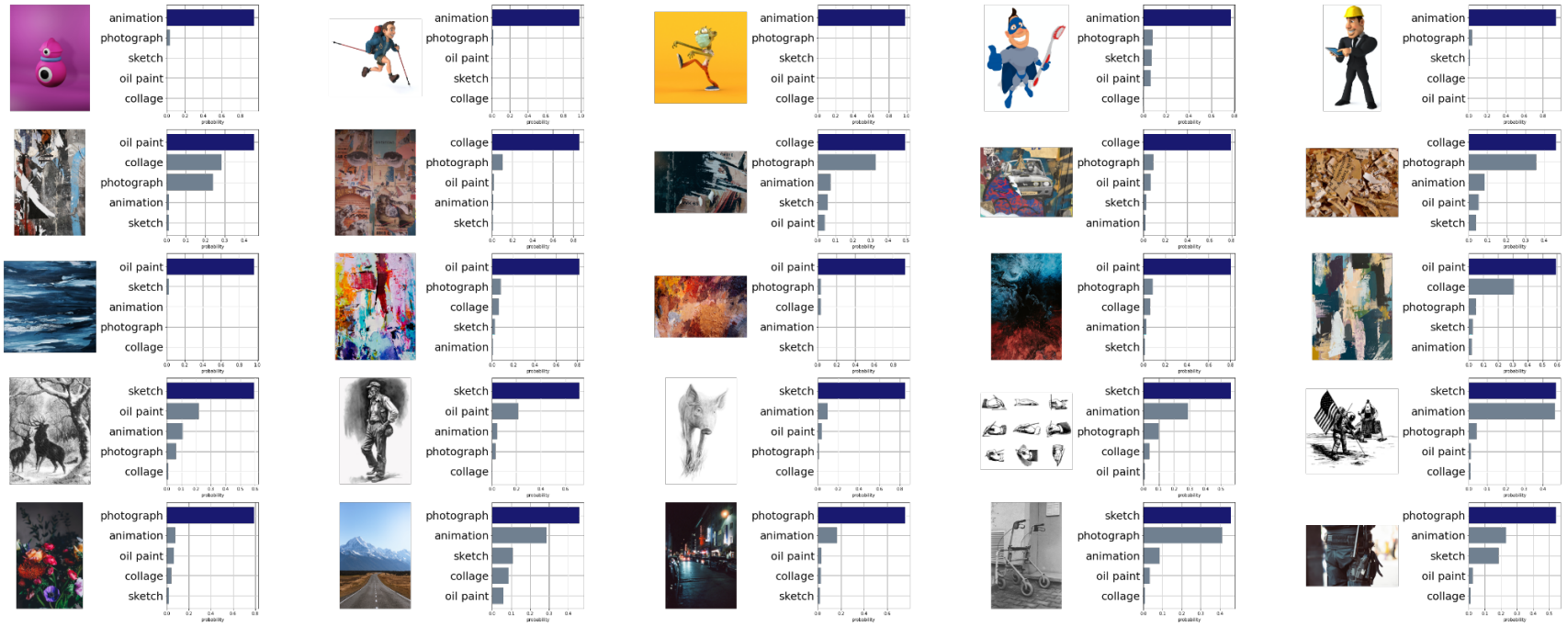


Figure 11: Result of CLIP's style recognition capacity test. CLIP successfully recognized each image style, proving its appropriateness on style measure. ("photograph" is an abbreviation for "photographic work".)

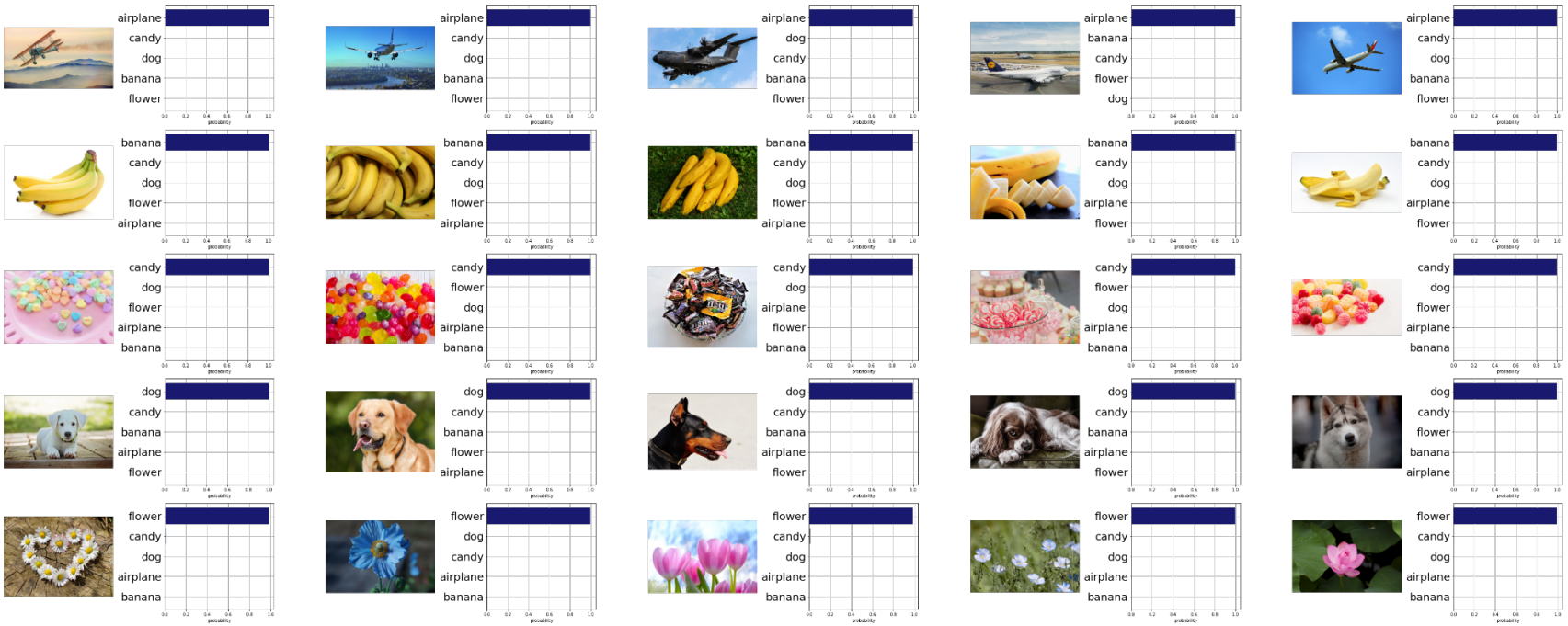


Figure 12: Result of CLIP's content recognition capacity test. CLIP confidently recognized each image content, proving its appropriateness on content measure.