PARASOL: Parametric Style Control for Diffusion Image Synthesis

Supplementary Material

1. Training Supervision Data

In order to obtain suitable data for training our model, a set of 500k triplets (output image x, content image y, style image s) is obtained through cross-modal search (See Section 3.1 for details). The process in which these images are obtained ensures a certain disentanglement between content in x and s and between style in x and y. These triplets are essential for the training of PARASOL and assist in disentangling the two attributes in which the network is conditioned. Fig. 14 shows a few examples of such triplets.

2. Baselines Comparison

Quantitative and qualitative evaluations are provided in the main paper (Section 4.2) comparing PARASOL to several image generation and neural style transfer (NST) methods.

The evaluations show how RDM [3] and ControlNet [53] are the methods that less accurately keep the fine-grained style and control details. RDM encodes both conditions using the same kind of encoding, without encouraging any disentanglement, leading to confusion of the network in which attributes should be transferred from each input condition. For the comparison, ControlNet was trained following the author's indications and using our set of triplets as training data. It only accepts content information given in textual format, so we use automatically generated captions from each content image y using BLIP [27]. Thus, the method was trained by feeding the generated captions as input and the style images as a conditioning that needs to be learnt. Their paper shows the method is able to learn a wide range of conditioning signals including sketches, segmentation maps and edgemaps. It doesn't, however, show any example in which the conditioning signal is not structure-based. Therefore, we hypothesize the architecture or parameters of ControlNet might not be suitable for successfully accommodating a condition such as style.

2.1. Comparison to eDiff-I

eDiff-I [2] is a diffusion-based method that generates images from text. It conditions the generation process on T5 text embeddings [33], CLIP image embeddings and CLIP text embeddings [32]. The use of CLIP image embeddings allows extending the method for style transfer from a reference style image.

This work wasn't included in the previous baseline comparison due to lack of open source code or public pretrained models. However, we show in Fig. 15 a visual comparison to a set of synthetic images they provide. Those images were generated by eDiff-I from long descriptive captions and the displayed style images. They also provide unstylized images generated from the same descriptions without conditioning on any style. eDiff-I incorporates T5 text embeddings in their pipeline, allowing it to process much more complex text prompts than those that can be encoded through CLIP. Therefore, in our comparison, we consider as content input the provided unstylized synthetic images generated from the same prompts.

3. Amazon Mechanical Turk Experiments

The results of 8 different Amazon Mechanical Turk (AMT) experiments were presented in the main paper, 6 of them comparing our method to different baselines in terms of style, content and overall preference (Section 4.2) and 2 comparing style and content interpolations to DiffuseIT [26] and RDM [3] (Section 4.5 (A)).

In particular, the instructions given to the workers in each task were the following:

- Image generation preference in terms of style: *"The photo has been transformed into the style of the artwork in multiple ways. Study the options and pick which most closely resembles the style of the artwork whilst also keeping the most structure detail in the photo."*
- Image generation preference in terms of content: *"The photo has been transformed into the style of the artwork in multiple ways. Study the options and pick which keeps the best structure of the content, from details in the content image."*
- Image generation overall preference: *"The photo has been transformed into the style of the artwork in multiple ways. Study the options and pick which most closely resembles the style of the artwork whilst also keeping the most structure detail in the photo."*
- Preferred method for content interpolation: *"These images have been generated by interpolating Photo1 and Photo2 and transferring the style of the Artwork. Study the options and pick which set of images better display a smooth transition from content/semantics of Photo1 and Photo2 while maintaining good image quality and a consistent style similar to the Artwork."*
- Preferred method for style interpolation: *"These images have been generated considering the structure in the Content image and an interpolation of styles from both Artworks. Study the options and pick which set of images better display a smooth transition from the style of Artwork1 to Artwork2 while maintaining good image quality and a consistent structure similar to the Content image."*

The first three instructions were used for separately comparing PARASOL to image generation methods and NST ones. A few examples of the images shown to the users in the AMT interpolation experiments are shown in Fig. 16, 17.

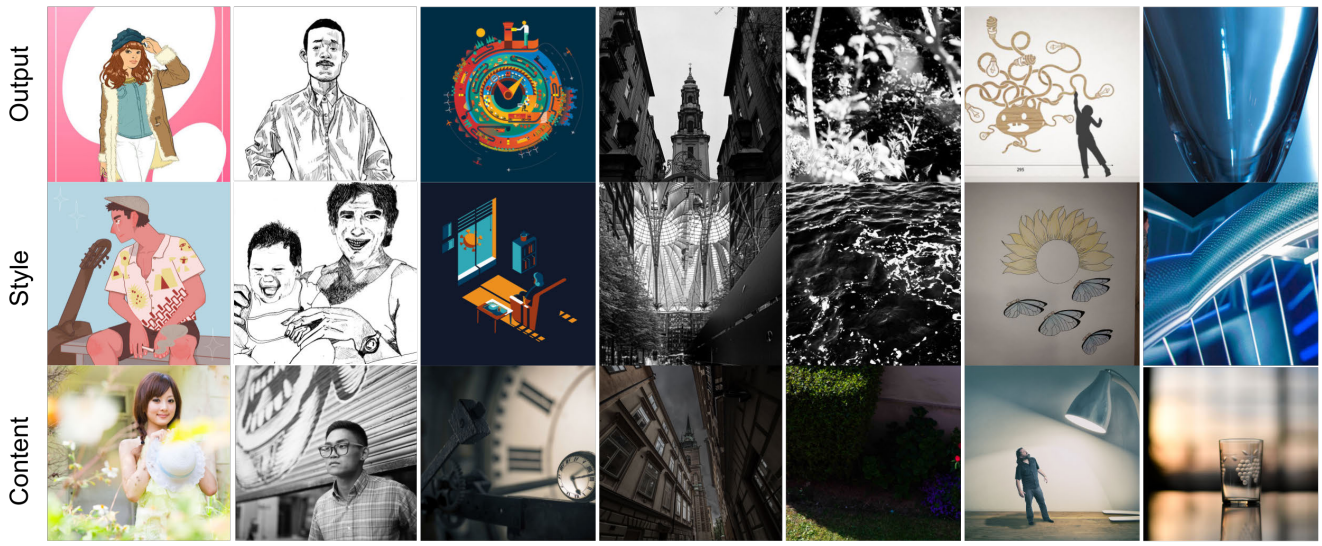

Figure 14. Visualization of triplets used for training our model. These triplets contain: output image x, similar image in terms of style s and similar image in terms of content y. During training, the method learns to reconstruct x by conditioning on y and s. The complete set of training triplets will be released as a second contribution.

Figure 15. Comparison to eDiff-I [2]. PARASOL closely transfers the fine-grained style of the style image while keeping the fine-grained details and structure of the content image. The concept of style, however, slightly differs between both works. While eDiff-I understands "style" as a collection of colours and structure in which the information should be presented, PARASOL focuses on the type of artistic style (e.g. oil painting, illustration...) and its fine-grained details such as types of brush strokes, while also transferring the overall colour tonalities.

4. Generative Search

Briefly introduced in Section 4.5 (B), generative search is presented as one of the main applications for our model. Fig. 18 shows a different example of how PARASOL can be used for either refining the search with a more fine-grained query in terms of style and semantics or for generating a synthetic image that closely matches the user's intent. As depicted in this example, style and/or content properties of different existent images can be combined for a more fine-grained search.

5. Additional Visualization Examples

We provide additional visualization examples for all experiments in Sections 4.4 and 4.5 (A), as well as an additional controllability experiment.

5.1. Images Generated from Textual Inputs

An example of images generated with PARASOL using textual vs. image inputs for style and/or content is provided in Fig. 19.

5.2. Images Generated with Different Lambda Values

PARASOL offers control in the amount of fine-grained content details that should be kept in the generated image vs. how much the image should be adapted to the new style. This can be controlled via the parameter λ (Fig. 20).

5.3. Images Generated with Different Classifier-Free Guidance Parameters

The classifier-free guidance parameters g_s and g_y indicate how much weight the style s and content y conditions should have in the generation of the new image. Fig. 21 visualizes the difference those values can make in the generated samples.

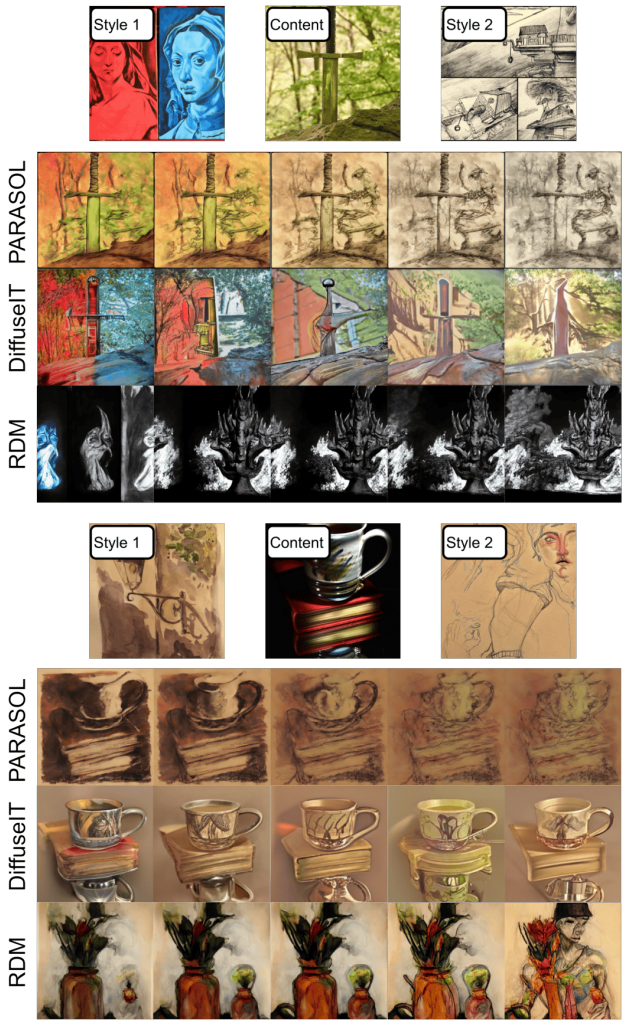

Figure 16. Baseline comparison for style interpolation. Visualization of images generated by PARASOL, DiffuseIT and RDM using interpolation between "Style 1" and "Style 2".

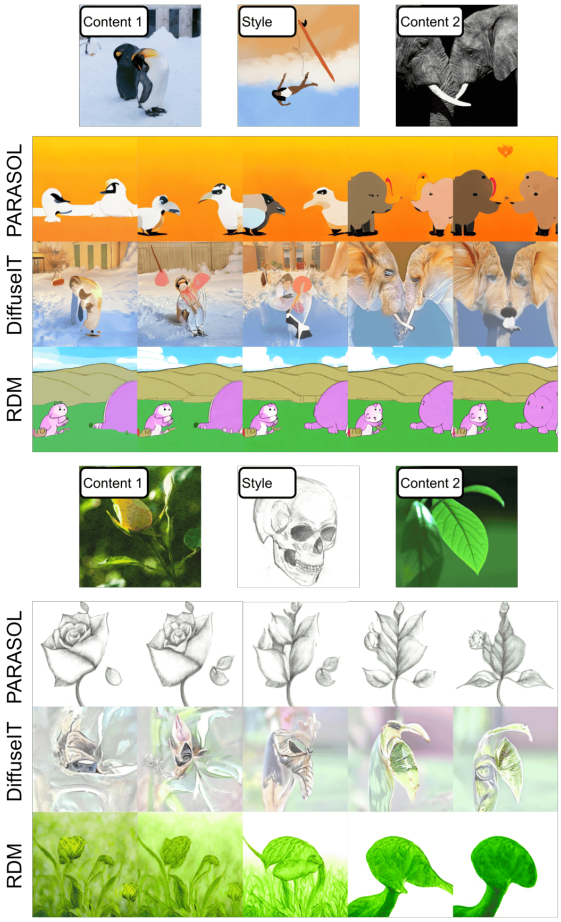

Figure 17. Baseline comparison for content interpolation. Visualization of images generated by PARASOL, DiffuseIT and RDM using interpolation between "Content 1" and "Content 2".

5.4. Images Generated by combining Different Lambda Values and Classifier-Free Guidance Parameters

Fig. 22 shows images generated by considering different pairs of g_s and λ values, while keeping g_y constant. The ratio of these two parameters defines how much creativity the model is allowed to introduce in the structure and stylization of the image.

5.5. Images Generated with Style and Content Interpolation

PARASOL allows the generation of images from a combination of different styles and/or contents (Fig. 23, 24). For combining the information from each pair of descriptors, their spherical interpolation is computed, considering a parameter $0 \leq \alpha \leq 1$. If $\alpha = 0$ only the first descriptor is considered, while $\alpha = 1$ indicates the second descriptor is the only one taken into account.

5.5.1 Style Interpolation

The use of a parametric model [37] for encoding the style condition allows the synthesis of images by interpolating different styles. The nuanced information this model is capable to encode allows the interpolation of very similar styles (Fig. 25) while also being able to interpolate more different styles (Fig. 26).

5.5.2 Content Interpolation

We encode the content information using CLIP [32] which is also a parametric model. Therefore, not only PARASOL can generate images by interpolating different styles, but it also allows the interpolation of different content information. The content information being interpolated can contain similar semantics (Fig. 27) or different ones (Fig. 28).

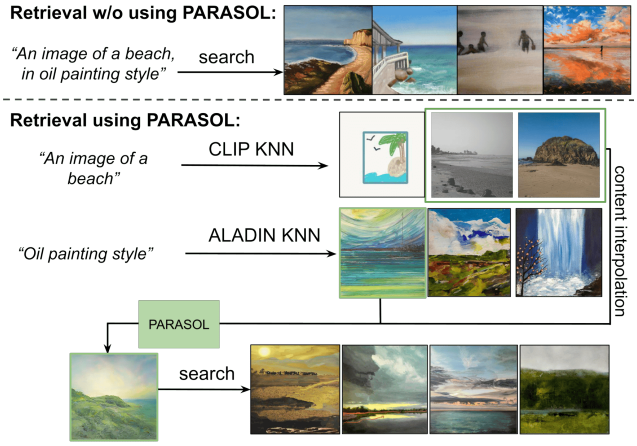

Figure 18. Use case for Generative Search. All PARASOL controllability tools, including interpolation capability, can be leveraged for obtaining a fine-grained query to refine the search and more closely match the user's intent.

Figure 19. Text vs. Image to describe style and content conditions. While providing an image to describe intended content or style provides more fine-grained details, textual inputs allow useful descriptions and unlimited creativity.

5.6. Images Generated with Different Fine-Grained Content Details

Section 4.4 (C) details how PARASOL can generate images with consistent semantics and fine-grained style while offering diversity in the fine-grained content details. Fig. 29 offers examples of this use case using $\lambda = 20$, $g_s = 5$ and $g_y = 5$. However, those parameters could be tuned for further control over the attributes of the final image.

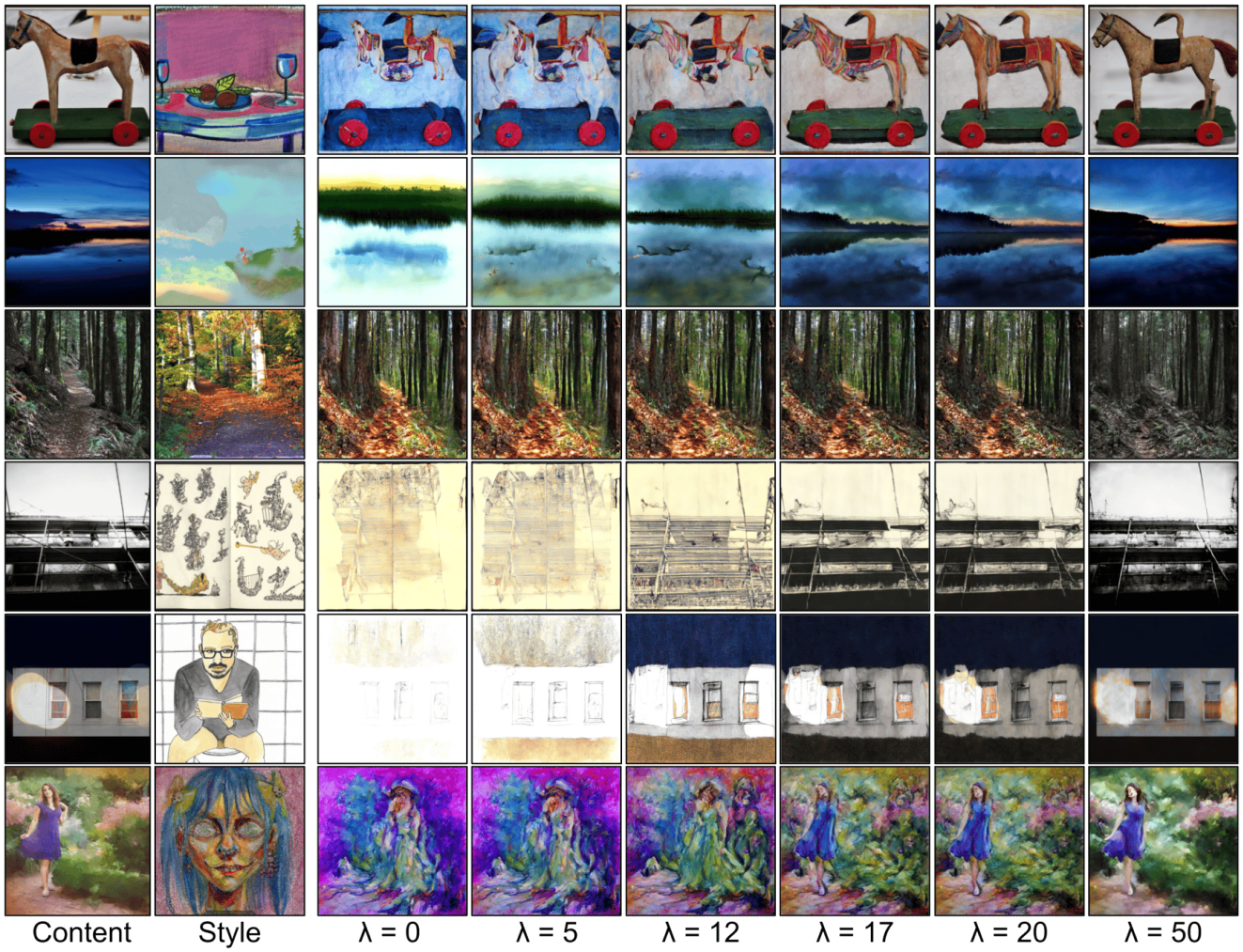

Figure 20. Images generated with different λ values. Higher values of lambda lead to more faithfulness in the structure and fine-grained details from the content input. Low lambda values lead to more stylised images that allow more creative structures and flexibility in fine-grained content details. In this example, $T = 50$, meaning lambda can take values from 0 to 50.

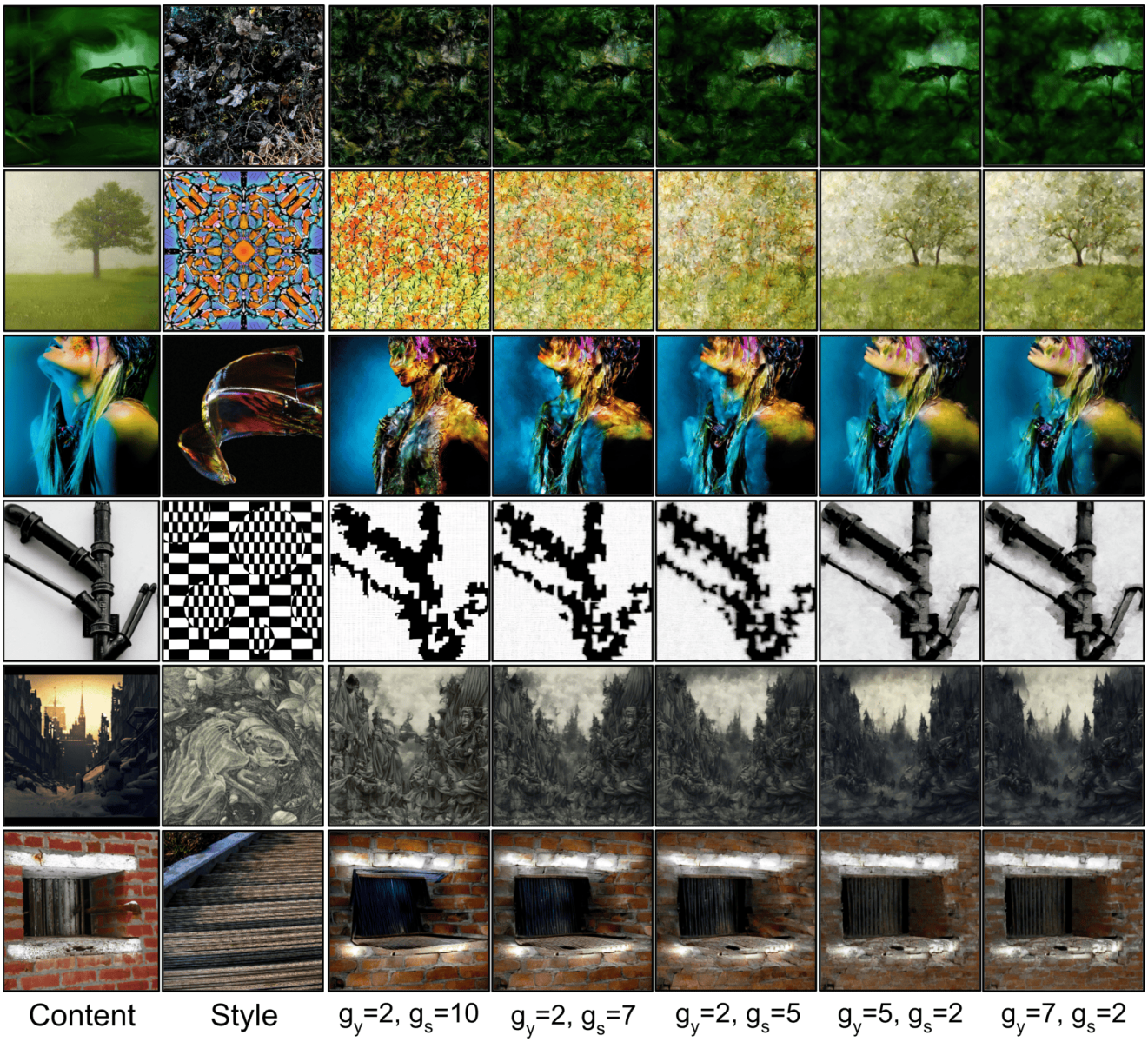

Figure 21. Images generated by PARASOL using different values for g_s and g_y. Fixing λ means that the structure of the content image is preserved in the same degree for all images. However, the balance between preserved semantics and style change with the ratio of both parameters.

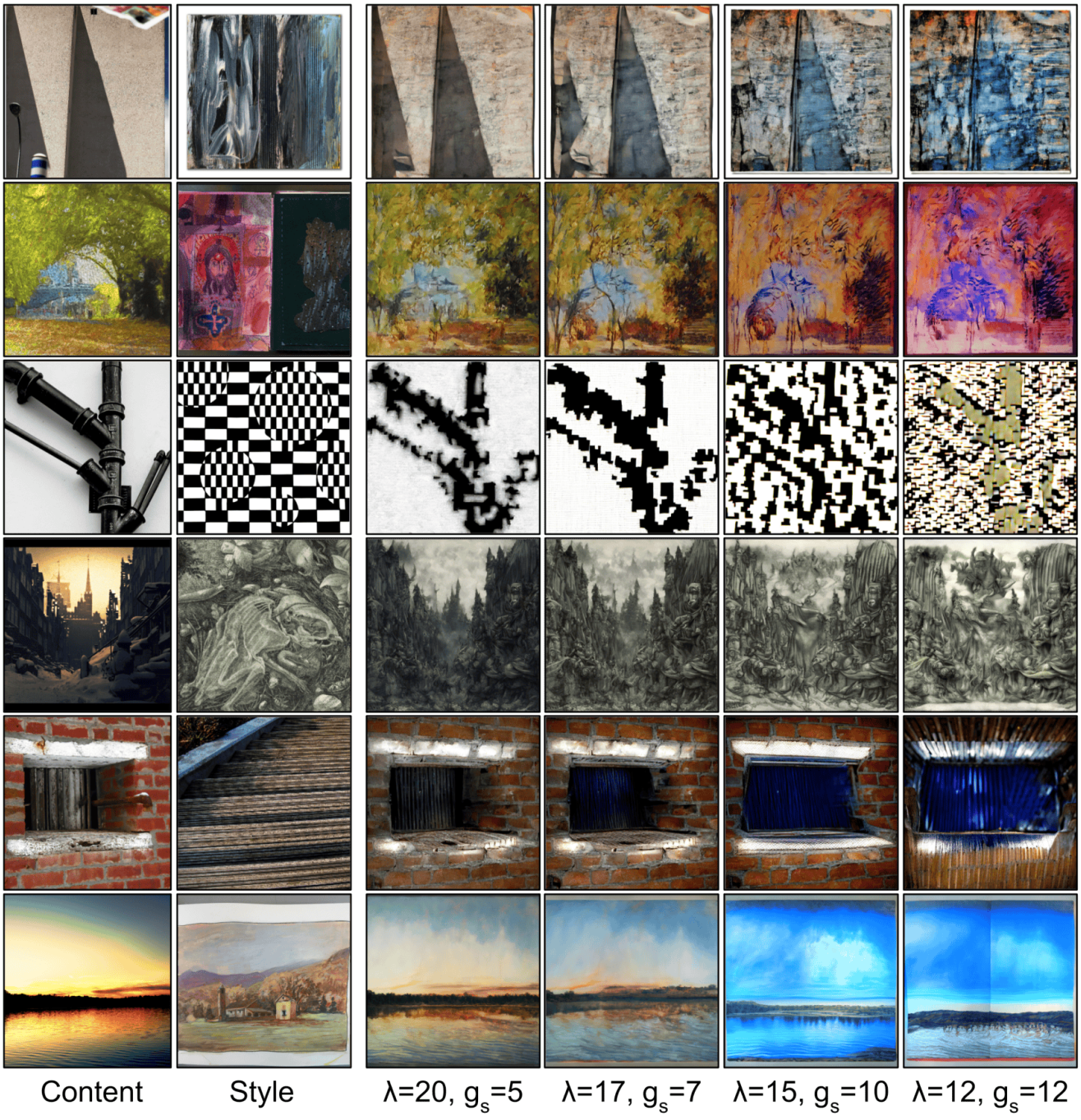

Figure 22. Images generated by PARASOL using different values for λ and g_s. High values of λ paired with low g_s lead to more faithful structures to the content input with more subtle stylization, while high values of g_s and low λ values lead to a more noticeable influence of the style image, with more space for creativity in terms of content details. The combination of both parameters allows for a wider range of options in terms of fine-grained controllability of the style and content details in the output image.

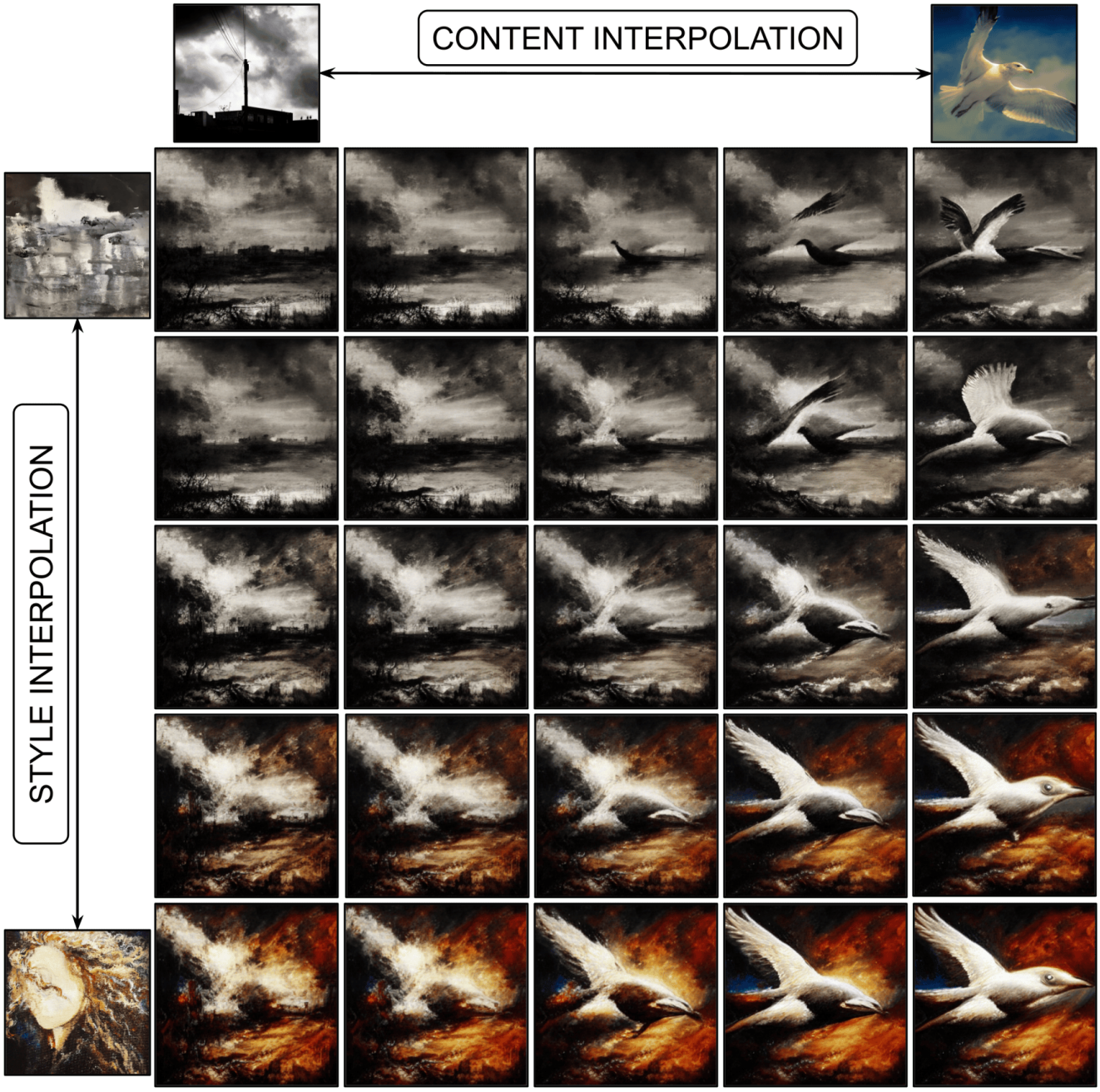

Figure 23. Style and Content interpolations. Visualization of different degrees of interpolation between two content images and two styles. For both style and content interpolations, values $\alpha = 0, 0.25, 0.5, 0.75, 1$ are considered.

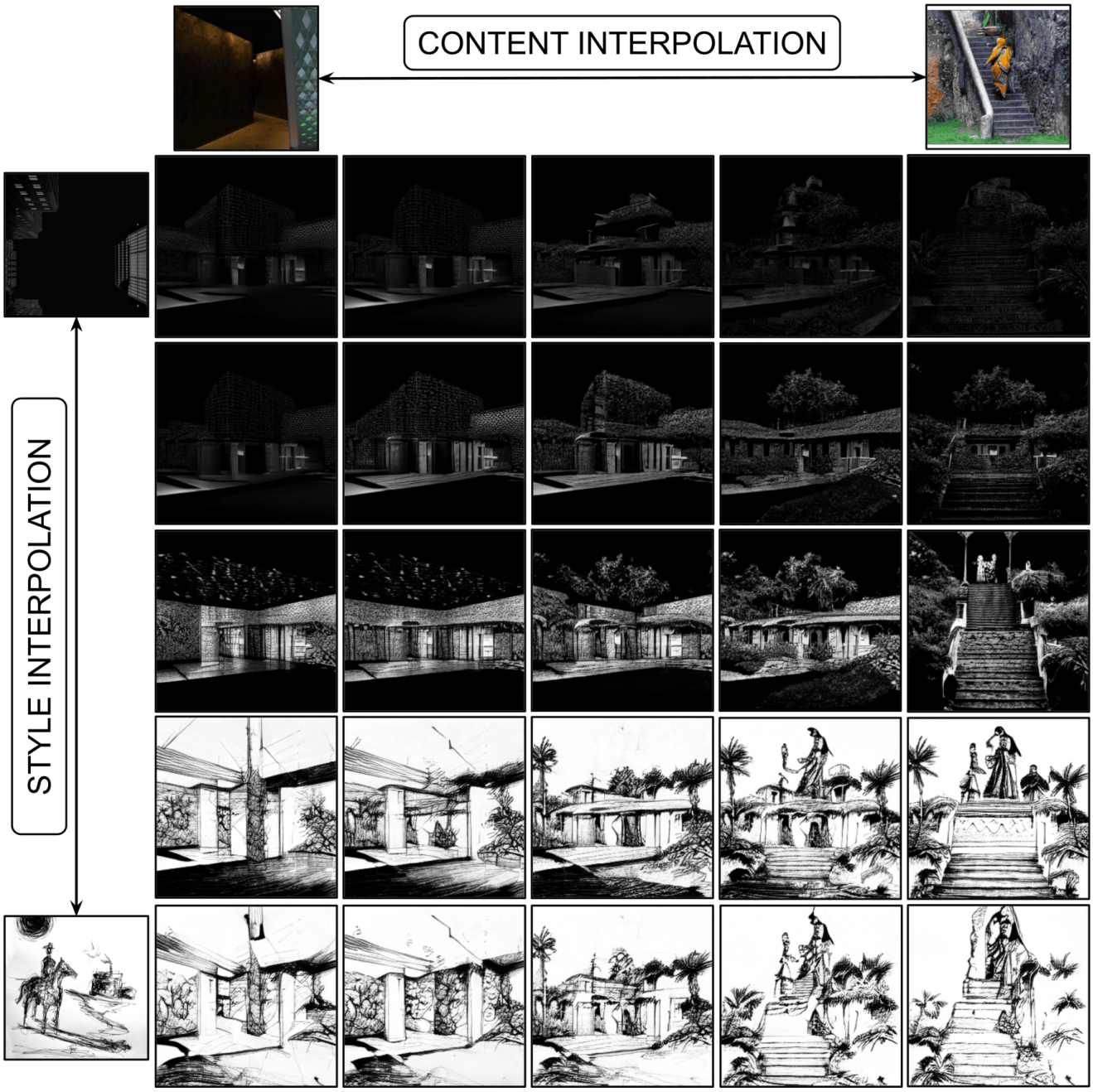

Figure 24. Style and Content interpolations. Second example of images generated by interpolating two content images and two styles in different degrees. For both style and content interpolations, values $\alpha = 0, 0.25, 0.5, 0.75, 1$ are considered.

Figure 25. Style interpolation with similar fine-grained styles. Images generated by conditioning on a content image and different interpolations of two very similar styles, transferring and combining their nuances and particular characteristics.

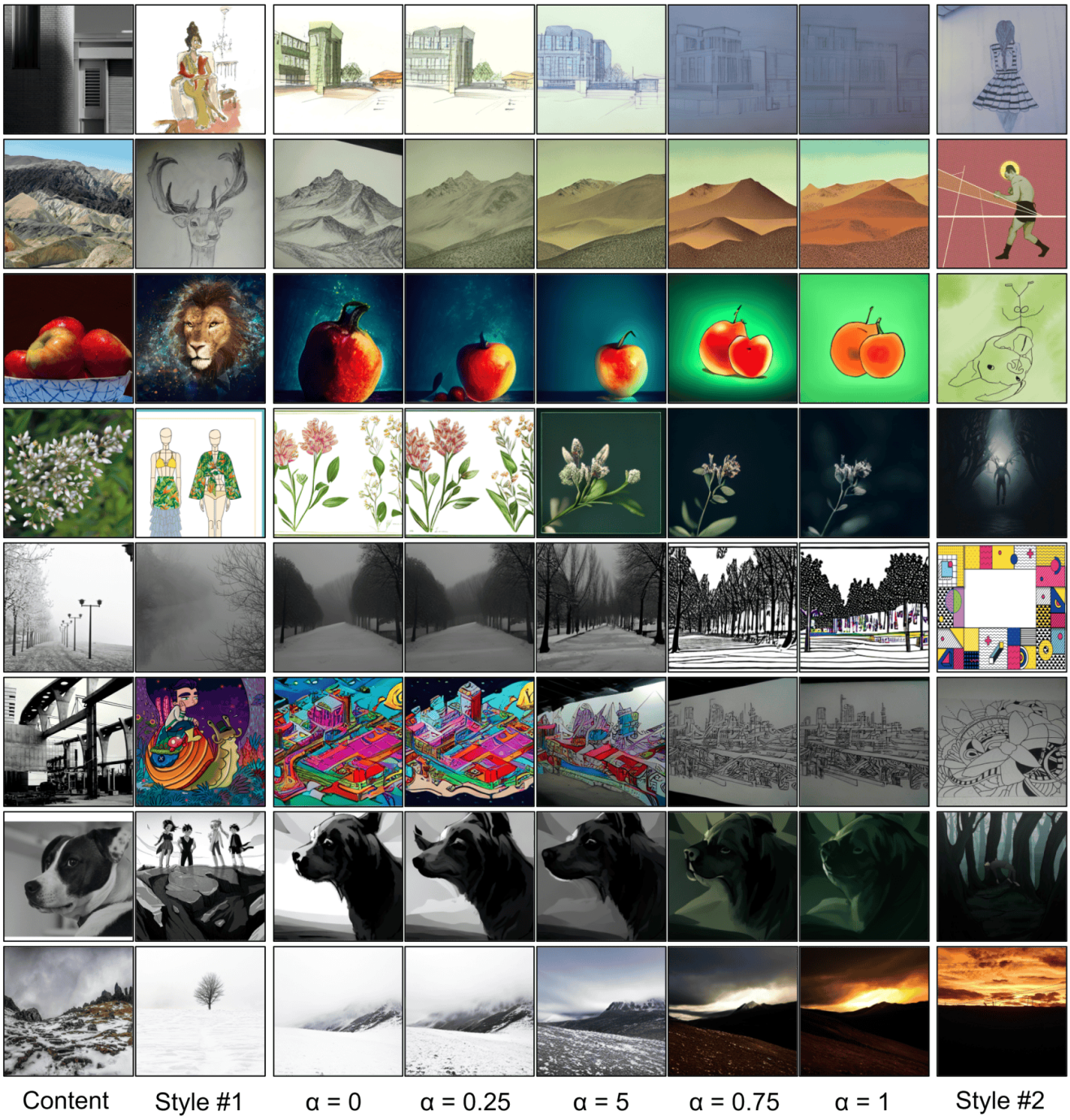

Figure 26. Style interpolation with visually different styles. Images generated by conditioning on a content image and different interpolations of two significantly different styles. PARASOL is able to smoothly transition between both styles, demonstrating its creative capabilities.

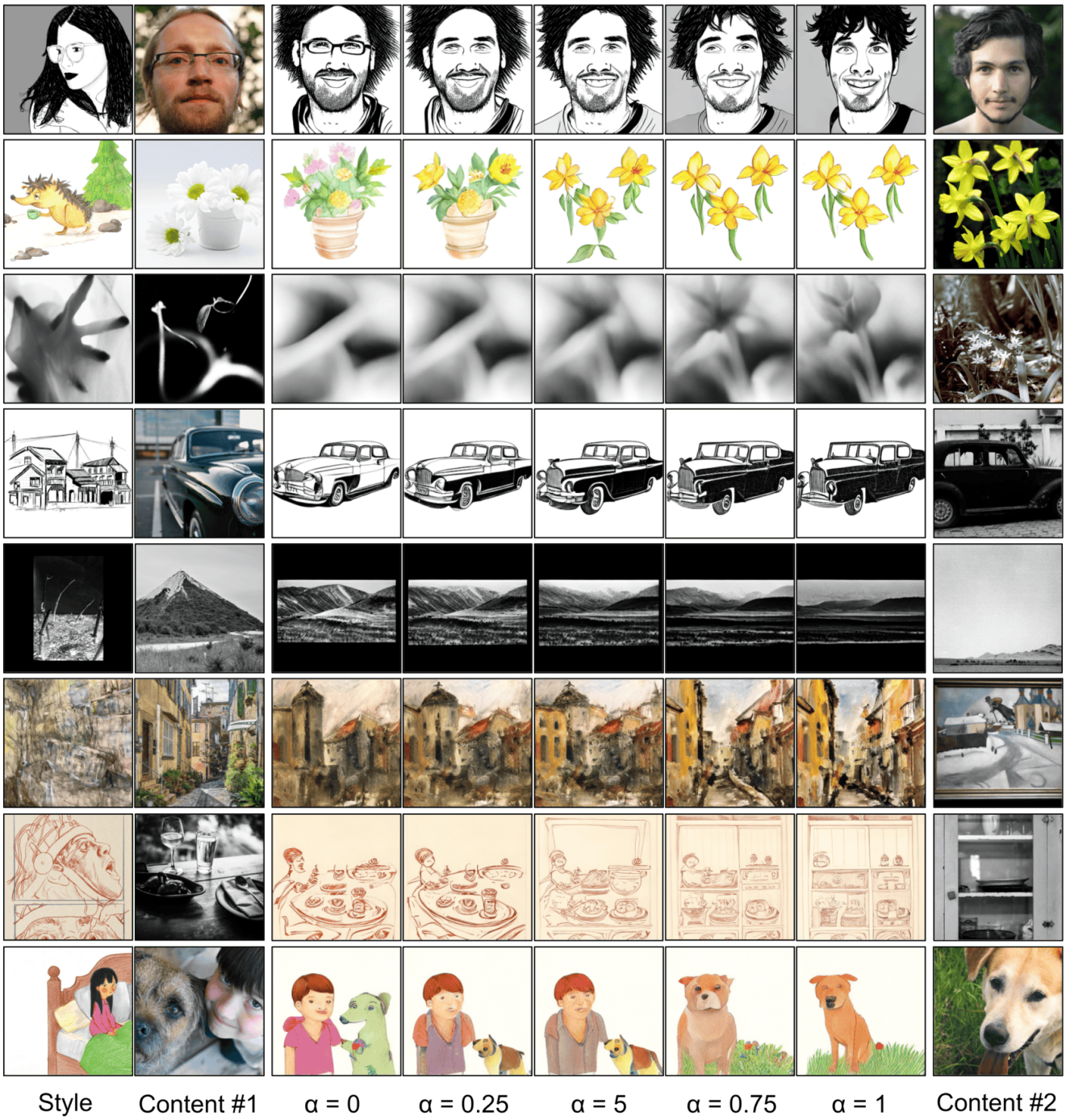

Figure 27. Content interpolation with similar semantics. Images generated by conditioning on a style and content information corresponding to the interpolation of two signals with similar semantics. PARASOL can capture the nuances and fine-grained details of each content input and combine them for generating brand new images.

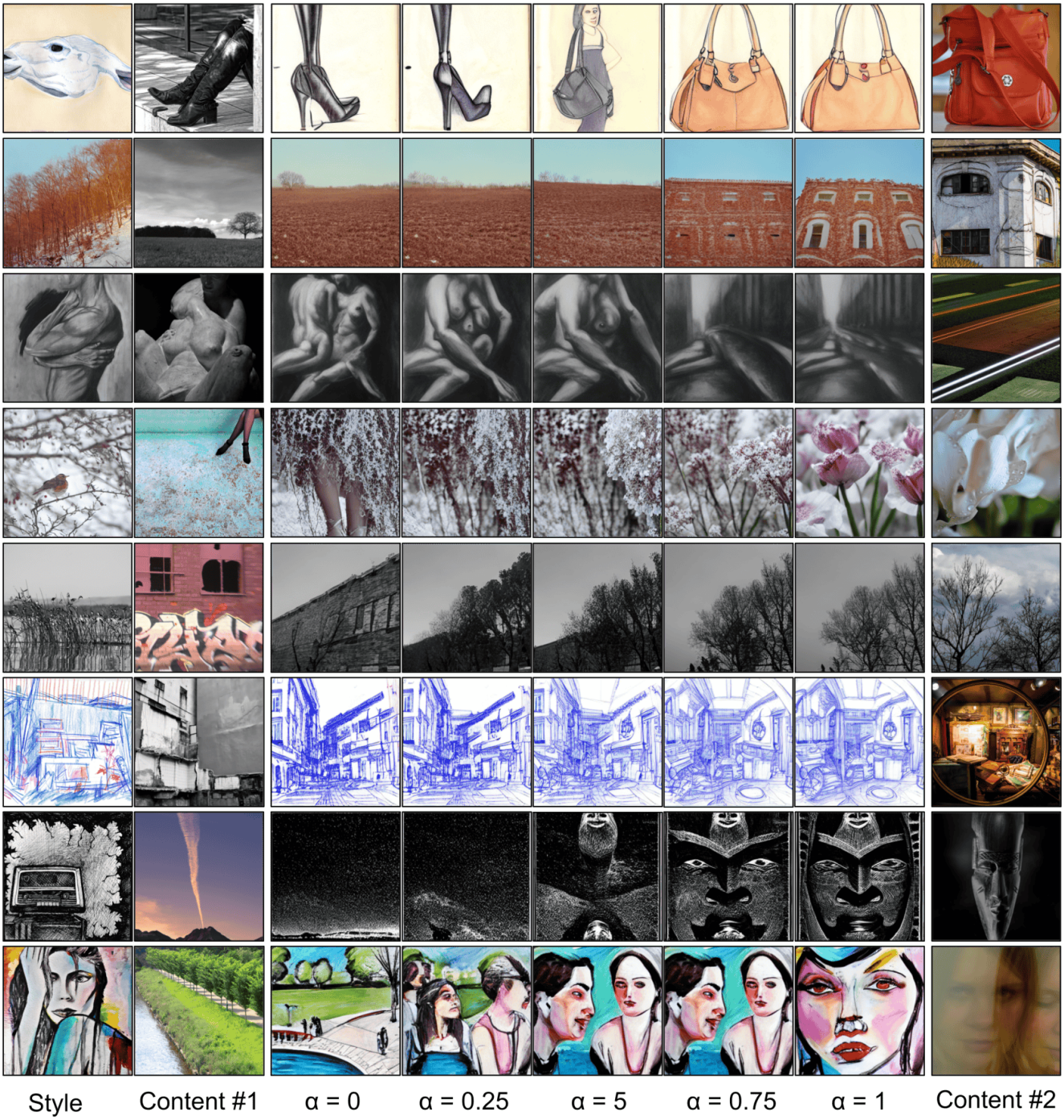

Figure 28. Content interpolation combining different semantics. Images generated by conditioning on a style and content information coming from the interpolation of two signals with different semantics. PARASOL can consistently transfer the style to both scenes or elements and combine them in a smooth way without losing image quality or realism.

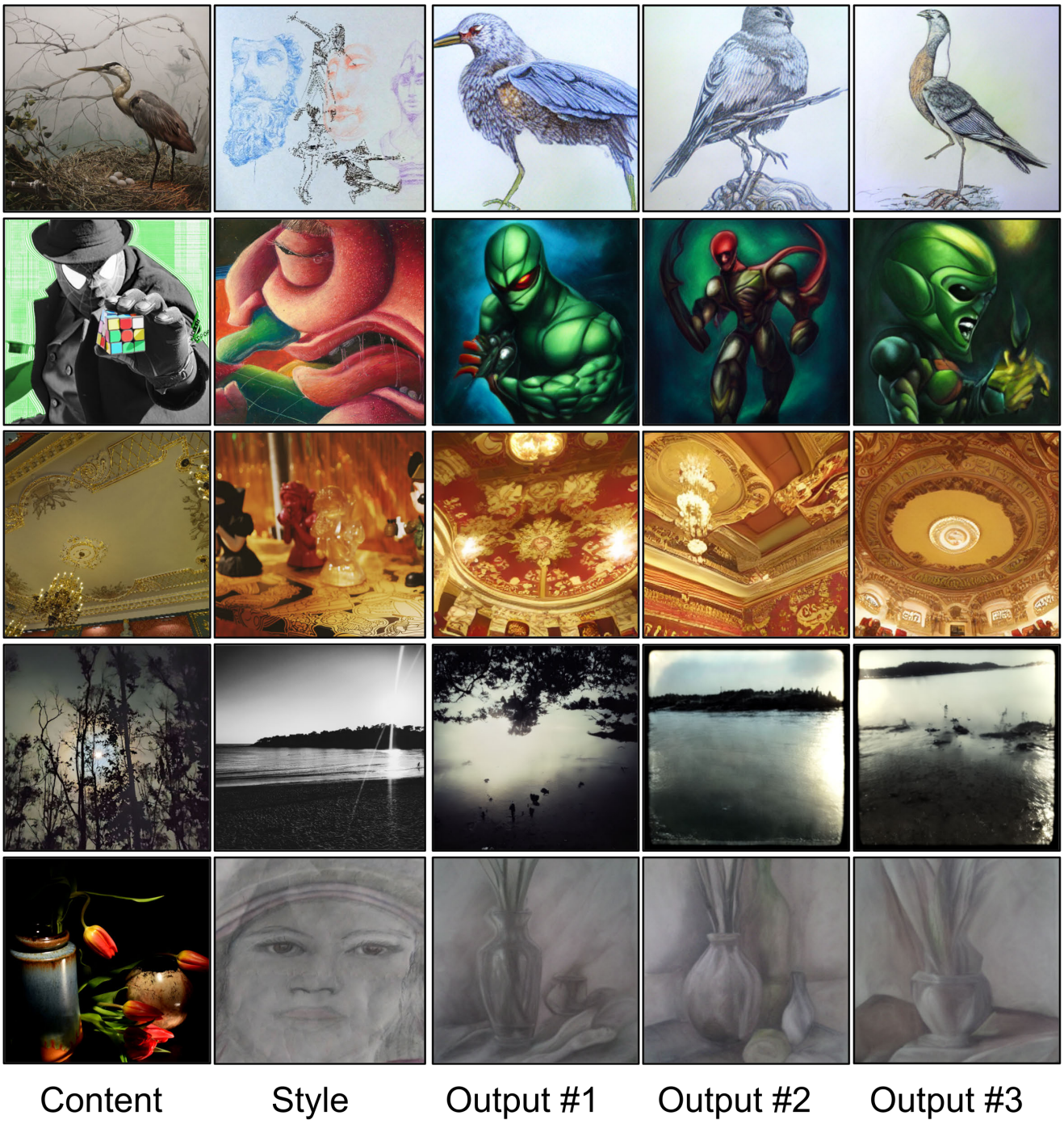

Figure 29. Diversity in fine-grained content. Images generated by PARASOL allowing flexibility in the fine-grained content details and image structure.