OmniStyle: Filtering High Quality Style Transfer Data at Scale Supplementary Materials

Ye Wang[1] Ruiqi Liu[1] Jiang Lin[2] Fei Liu[3] Zili Yi[2] Yilin Wang[4*] Rui Ma[1,5*]
[1]Jilin University [2]Nanjing University [3]ByteDance [4]Adobe
[5]Engineering Research Center of Knowledge-Driven Human-Machine Intelligence, MOE, China

Figure 1. Examples of content images across 20 categories. Each image is randomly selected from its respective category, showcasing the diversity of the categories and the high quality of the generated images.

1. More Details about OmniStyle

In this section, we provide more details about OmniStyle. Section 1.1 delves into the process of generating content images, while Section 1.2 elaborates on the specific details of the collected style images. More triplet examples from our dataset are presented in Section 1.3.

1.1. Content Images Generation

We adopt a generative approach to obtain content images, which offers the advantages of effectively avoiding issues related to copyright and privacy while ensuring high-quality data. Specifically, we first define 20 common image categories For each category, we utilize ChatGPT [2] to automatically generate 100 textual descriptions. Based on these descriptions, we employ FLUX [1] to generate the corresponding images, resulting in a total of 2,000 content images across the 20 categories. Figure 1 provides an overview of all 20 categories and the exemplar generated content images. Each image is rendered at a 1K resolution, guaranteeing high visual clarity and fidelity. Table 1 presents the some exemplar textual prompts for generating the animal category as a reference.

Table 1. Ten textual prompts for the animal category.

1. A swan swimming gracefully on a calm lake at sunset.
2. A leopard resting on the branch of a tree in the jungle.
3. A close-up of a ladybug crawling on a green leaf.
4. A family of deer grazing in a peaceful meadow.
5. A family of ducks swimming in a calm pond.
6. A polar bear swimming in the icy waters of the Arctic.
7. A family of ducks swimming in a calm pond.
8. A majestic lion standing on a rocky ledge at sunset.
9. A hawk swooping down to catch its prey.
10. A close-up of a spider spinning a web.

1.2. Style Images Collections

Existing style image datasets typically focus on a limited range of traditional style categories, such as oil painting, watercolor, and sketches. In contrast, our study highlights a broader and more fine-grained spectrum of style categories, encompassing up to 1,000 distinct styles collected from the Style30K [7] dataset. In Figure 2, we provide examples of some style images for reference.

1.3. Examples of Triplet

We provide additional examples of triplet from the OmniStyle-1M dataset, as illustrated in Figure 3. Both the content images and the generated stylized images used in this study maintain a resolution of 1024*1024,

*Corresponding authors.

Figure 2. A subset of style images in OmniStyle Dataset.

Figure 3. Examples of Triplet from the OmniStyle-1M dataset. From left to right, each set of images is arranged as: content image, style image, and resulting stylized image.

ensuring high-quality data. Moreover, it can be found the stylized result images exhibit a high degree of style consistency with the designated reference styles, which reflects the superiority of our constructed dataset in preserving style.

2. More Details about OmniFilter

In this section, we provide more details about the evaluation schemes for content preservation, style consistency and aesthetic appeal, which are the three key aspects of the OmniFilter.

Evaluation on Content Preservation. We present the detailed model architecture for content preservation evaluation as shown in Figure 4.a. In this architecture, CLIP [10] is employed to compute the similarity score between the stylized image and the content text, serving as a measure of semantic similarity. Simultaneously, DINOV2 [9] is used to calculate the structural similarity between the stylized image and the original content image. This process does not require any training or fine-tuning. Specifically, the pre-trained weights for CLIP [10] are from "clip-vit-large-patch14," and for DINOV2 [9], "dinov2-large" is utilized.

Evaluation on Style Consistency. As illustrated in Figures 4.b and c, we present the training and inference architecture for style consistency evaluation. In this architecture, we adopt a contrastive learning approach, leveraging the Style30K [7] dataset to construct positive and negative sample pairs for updating the CLIP [10] im-