UnZipLoRA: Separating Content and Style from a Single Image

Chang Liu, Viraj Shah*, Aiyu Cui†, Svetlana Lazebnik
University of Illinois, Urbana-Champaign
changl25, vjshah3, aiyucui2, slazebni@illinois.edu

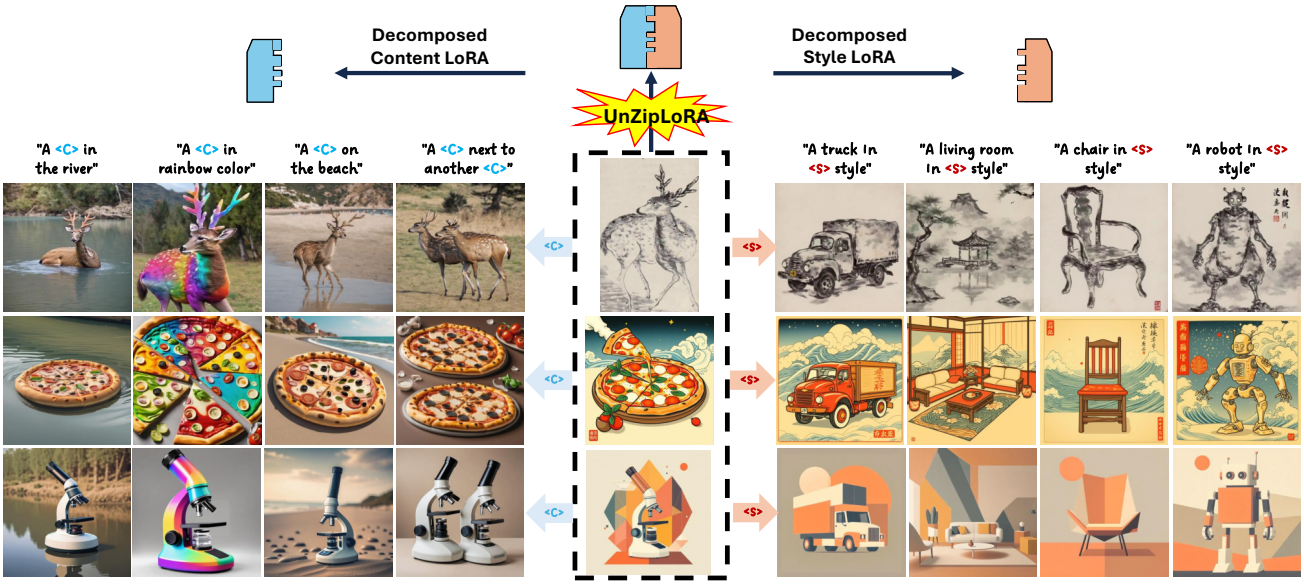

Figure 1. Given a single input image (middle), **UnZipLoRA** learns a disentangled *content* or *subject* LoRA (left) and *style* LoRA (right) that can be used to generate new images with the learned concepts.

Abstract

This paper introduces UnZipLoRA, a method for decomposing an image into its constituent subject and style, represented as two distinct LoRAs (Low-Rank Adaptations). Unlike existing techniques that focus on either subject or style in isolation, or require separate training sets for each, UnZipLoRA disentangles these elements from a single image by training both the LoRAs simultaneously. UnZipLoRA ensures that the resulting LoRAs are compatible, i.e.*, can be seamlessly combined using direct addition. UnZipLoRA enables independent manipulation and recontextualization of subject and style – generating variations of each, applying the extracted style to new subjects, and recombining them to reconstruct the original image or create novel variations. To address the challenge of subject and style entanglement, UnZipLoRA employs a novel prompt separation technique, as well as column and block separation strategies to accurately preserve the characteristics of subject and style and ensure compatibility between the learned LoRAs. Evaluation with human studies and quantitative metrics demonstrates UnZipLoRA's effectiveness compared to other state-of-the-art methods, including DreamBooth-LoRA, Inspiration Tree, and B-LoRA.*

1. Introduction

Imagine an artist inspired by a single image – perhaps a microscope rendered in a unique illustration technique (Fig. 1). The image depicts both the *content* or *subject* (the microscope with its specific shape and details) and *style* (the characteristic artistic technique). What if we could disentangle these intertwined elements, enabling the artist to extract and manipulate them independently? This capability, long sought in the computer vision literature [36], would open up new avenues for artistic expression, allowing for the destylization and recontextualization of the subject in different conditions, the application of the extracted style to new subjects, and the creation of entirely novel combina-

* Currently at Google. † Currently at Amazon.
Project page: https://unziplora.github.io

tions while preserving both the original subject and style.

The recent surge of diffusion models [13, 28, 35] has unlocked previously unprecedented automatic image generation capabilities. The primary means of controlling such models is through text prompts, but text-based conditioning is inadequate for capturing the details of nuanced concepts, such as specific object instances (my dog) or individual styles (my child's crayon drawing). Example-driven generation is highly desired in such scenarios. To this end, model fine-tuning methods like DreamBooth [29] and StyleDrop [33] capture subject or style from reference image(s) in a way that allows for novel renditions. However, such approaches tend to focus on either content or style in isolation and cannot be easily made to capture both or perform disentanglement. On the other hand, stylization techniques like ZipLoRA [32] and B-LoRA [6] can combine subject and style, but require the training of two separate models using subject- and style-specific input images. However, the scenario we consider requires the opposite – decomposing the subject and style from a single image.

In this work, we introduce **UnZipLoRA**, a novel method that deconstructs an image into its constituent subject and style, represented as two distinct LoRAs (Low-Rank Adaptations [14]) trained simultaneously. These LoRAs can be used independently to generate variations of the subject or style and allow for recontextualizations. Moreover, our joint training method ensures that the resulting LoRAs are inherently compatible, *i.e.*, can be seamlessly combined by direct addition to reconstruct the original image or to generate novel compositions of subject and style while preserving their fidelity. In fact our approach can be seen as a next stage of *concept extraction* – a problem previously studied in concept decomposition methods like Inspiration Tree [37] and CusConcept [40] that rely on textual inversion [8] to learn multiple text embeddings corresponding to the *subconcepts* within a set of images. However, textual inversion alone, without fine-tuning of weights, does not provide adequate control over or fidelity of the extracted concepts, which tend to remain generic and fail to capture nuances of the input object/style.

As suggested by the name, UnZipLoRA operates in the opposite direction of ZipLoRA [32], which is focused on merging independently trained subject and style LoRAs. While ZipLoRA addresses the challenge of combining pre-existing LoRAs, UnZipLoRA tackles the inverse problem: disentangling a single image into its subject and style components such that the resulting LoRAs can be used either together or separately. Mathematically speaking, the decomposition problem is ill-posed and cannot be trivially derived from the approach of ZipLoRA [32].

Our key challenge is to learn two independent LoRAs simultaneously using only a single input image as supervision while ensuring that the resulting LoRAs correctly capture the subject and style concepts. Typical LoRA fine-tuning operates by binding the LoRA weights with the trigger phrase representing a specific subject or style concept in the input image – such as "a $<c>$ in $<s>$ style". If we apply such a naive approach in our case, the presence of both the trigger phrases $<c>$ and $<s>$ in a single input prompt makes it difficult for the subject and style LoRAs to bind to the correct trigger phrase, resulting in cross-contamination. To solve this problem, we propose a novel *prompt separation* strategy that uses different prompts for each LoRA and the base model, and then combines them together in the intermediate feature space of the diffusion model in such a way that the loss for each LoRA can be calculated jointly using only the input image as supervision.

While prompt separation allows for joint training of LoRAs, the resulting LoRAs may not be compatible with each other, *i.e.*, combining them through direct addition may produce poor quality recontextualizations. To make them inherently compatible, we also propose *column separation* and *block separation* strategies. In particular, *column separation* determines the importance of each column of LoRA weight matrix and adaptively assigns each column to either a subject LoRA or style LoRA using a dynamic importance re-calibration strategy. Such disjoint assignment ensures that high-importance columns from each LoRA remain decoupled. *Block separation* reserves some blocks of the U-net predominantly for style or for the subject, providing a further degree of disentanglement.

We demonstrate the effectiveness of UnZipLoRA for accurate separation of content and style using both human studies and quantitative metrics. Our results show a clear advantage of our method over separate LoRA fine-tuning via DreamBooth [29], concept separation via Inspiration Tree [37], and even the most recent state-of-the-art B-LoRA method [6]. We also showcase our method's ability to preserve the concept fidelity for a wide array of recontextualizations – whether for using subject or style separately or together. In addition, our method provides the valuable capability for cross-composition of subject and style LoRAs obtained from different images. Finally, we demonstrate generalizability of our approach by providing results on KOALA – a newer, more efficient text-to-image model [19].

2. Related Work

Fine-tuning diffusion models is an effective way to personalize the text-to-image (T2I) models to depict specific concepts based on textual descriptions. Textual Inversion [8] optimizes the text embedding to represent a specific visual concept. DreamBooth [29] fine-tunes the diffusion model itself to better capture an input concept from a small number of images, while another group of methods [10, 17] aim to optimize specific parts of the networks to capture visual concepts. Most personalization approaches

have quickly adopted LoRA [14], a fine-tuning technique that only optimizes a small subset of weights by low-rank approximations, as it is efficient for training and can mitigate overfitting problems.

Image stylization is an area of research dating back at least 20 years [5, 12]. Great advances in arbitrary style transfer were achieved by approaches based on convolutional neural networks [9, 15, 16, 20, 24]. With the advent of deep generative models, a variety of approaches have attempted to fine-tune a pre-trained Generative Adversarial Network (GAN) or diffusion model for stylization [3, 4, 7, 18, 21, 23, 31, 33, 38, 41–43]. While these works provide valuable insights into style learning using generative models, the task we attempt to solve is the opposite: instead of stylizing a content image, we attempt to decompose a stylized image into its subject and style.

Content-style decomposition. The task of subject-style decomposition can also be seen as a type of *concept extraction*. Inspiration Tree [37] aims to learn multiple embeddings corresponding to hierarchical subconcepts within a set of images. However, its reliance on textual inversion [8] limits this method to primarily producing text embeddings rather than the full LoRA weights needed for granular control over generation. Similarly, CusConcept [40] decomposes an image into visual concepts by learning customized embeddings. However, it does not produce dedicated LoRAs for each concept. Its reliance on computationally expensive LLMs and lack of explicit content-style separation limits its practical applicability. U-VAP [39] is a fine-tuning method that allows users to specify desired attributes from a set of images, enabling the disentangled use of visual concepts in diverse settings. However, this method requires elaborate data augmentation aided by LLMs, posing scalability challenges. ConceptExpress [11] explores unsupervised concept extraction and recreation by leveraging the inherent capabilities of pre-trained diffusion models. However, its reliance on localized masks extracted from the U-Net limits its application to concepts that can be localized within an image. This constraint prevents it from effectively handling global concepts like style. This limitation applies to Break-A-Scene [1] too, which relies on localized masks.

The method most directly aimed at content and style separation is B-LoRA [6]. The authors of B-LoRA analyze the architecture of the SDXL base model [26] to find U-Net blocks most responsible for capturing content and style. By *independently* training content and style LoRAs restricted to the respective blocks, they obtain models that can be successfully mixed for various stylization applications. By contrast, UnZipLoRA trains both the LoRAs *simultaneously*. In our attempt to take advantage of B-LoRA's block constraints for subject-style separation, we find that they are coarse-grained and thus insufficient to disentangle the subject in joint training. Therefore, we extend the block separation to more blocks and perform more fine-grained disentanglement. Together with our improved block separation strategy, the prompt separation strategy of UnZipLoRA provides a significant improvement in disentanglement ability, while column separation further enhances the fidelity of fine details.

3. Method

3.1. Preliminaries

Diffusion models [13, 28, 35] are state-of-the-art deep generative models renowned for their ability to synthesize high-quality photorealistic images. In this work, we focus on Stable Diffusion XL (SDXL) [26], a widely used U-Net-based latent diffusion model (LDM) [28] known for its strong performance. In Section 4.5, we also show results on the newer, more efficient KOALA model [19].

Low-Rank Adaptation (LoRA) is an efficient fine-tuning method for adapting large vision or language models to new tasks [14, 30]. LoRA assumes that weight updates ΔW during fine-tuning have a low intrinsic rank and can thus be decomposed into two low-rank matrices, $B \in \mathbb{R}^{m \times r}$ and $A \in \mathbb{R}^{r \times n}$, such that $\Delta W = BA$, where r is the intrinsic rank of ΔW with $r << \min(m, n)$. During training, only A and B are updated while keeping the original weights W_0 constant. The updated weights become $W = W_0 + BA$. In case of text-to-image LDM models, model customization can be achieved by using LoRA fine-tuning to minimize reconstruction loss $\mathcal{L}_{DB}$ as proposed by DreamBooth [29].

3.2. Problem Setup

We aim to extract content and style from a single input image by learning two distinct LoRAs simultaneously. Given a pre-trained diffusion model with weights $\{W_0^i\}$, we learn two models: content LoRA $L_c = \{\Delta W_c^i\}$ and style LoRA $L_s = \{\Delta W_s^i\}$. Here, i denotes the index of layers of the diffusion U-Net. In the following, we will skip the superscript i as we operate over all LoRA-enabled weights of the base model. Once trained, the resulting LoRAs can be used either separately or together to achieve various recontextualizations as depicted in Fig. 2.

We follow the standard prompt construction strategy "a $<c>$ in $<s>$ style" with trigger phrases $<c>$ and $<s>$ to describe the content and style respectively [6, 17, 29, 32, 33]. Following prior works such as DreamBooth [29], the subject trigger phrase $<c>$ is formed with a unique token followed by the subject class label (e.g., 'sks dog'), and the style trigger phrase $<s>$ consists of a generic description of style such as 'watercolor painting' [6, 32, 33]. We find that a single-word class label for the subject, and a high-level, brief (2-3 word) description for style are sufficient to effectively guide our method (see the supplementary material for further discussion of trigger phrase selection).

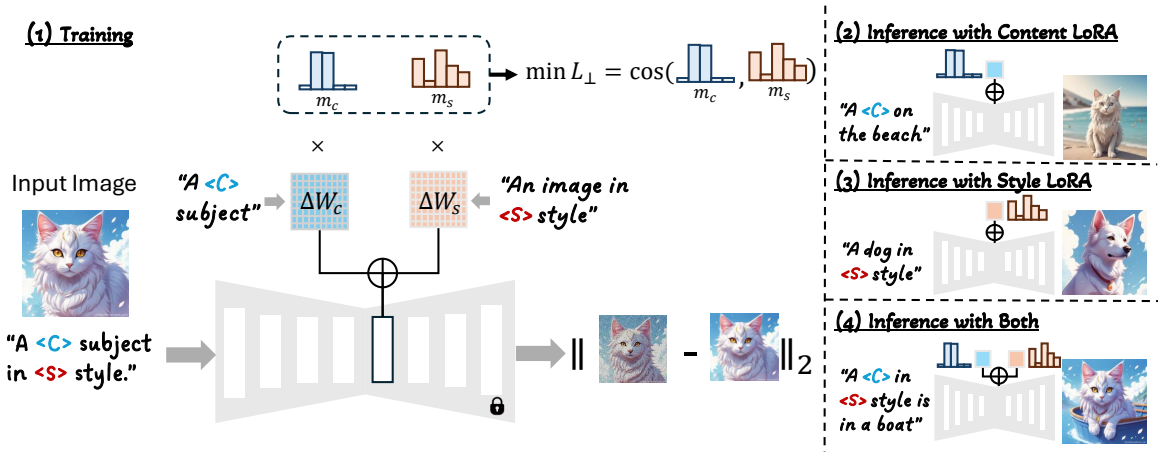

Figure 2. **Method Overview.** (1): Our training pipeline takes three prompts to learn two separate LoRAs. The column masks are introduced to establish the orthogonality. (2)-(4): Inference with individual learned LoRA or the combined LoRA.

3.3. UnZipLoRA

In this section, we present our UnZipLoRA approach, which relies on three key components to ensure accurate disentanglement. First, **prompt separation** (Sec. 3.3.1) allows for training the subject and style LoRAs simultaneously. Second, **column separation** (Sec. 3.3.2) uses a dynamic importance recalibration strategy to make the columns of resulting LoRAs mutually orthogonal. Third, we introduce **block separation** (Sec. 3.3.3) that reserves certain style/subject-sensitive blocks of the SDXL U-Net only for style/subject learning to improve the fine details.

3.3.1. Prompt Separation

Our core challenge lies in the utilization of a single image as supervision for training two separate LoRAs. Since the input image contains both subject and style, we must use both LoRAs L_c and L_s together to calculate the loss during training. If we also use $<c>$ and $<s>$ in the input prompt, as in "A $<c>$ in $<s>$ style", the training would lead to cross-contamination since the tokens corresponding to the style descriptor $<s>$ would be attended to by the cross-attention layers of the subject LoRA, and vice versa. This is evident in the Baseline (DreamBooth-LoRA) row of Fig. 5, where both the subject and style LoRAs are overfitted to input image.

Cross-attention layers in diffusion models are responsible for learning the text conditioning, and play a crucial role in binding the target concepts to the corresponding parts of the prompt. In a typical cross-attention layer of the diffusion U-Net, the prompt embedding x is mapped to keys K and values V in the transformer using weights W_0:

$$K(x) \text{ or } V(x) = W_0^T x\,. \tag{1}$$

If we add the content and style LoRAs to the base model and use both $<c>$ and $<s>$ in the prompt x, the mapping in the cross-attention layer becomes

$$K(x) \text{ or } V(x) = (W_0 + \Delta W_c + \Delta W_s)^T x\,. \tag{2}$$

This allows the content (resp. style) LoRA to attend to style (resp. content) tokens, resulting in cross-contamination. Instead of the naive strategy in eq. (2), we propose to calculate three sets of keys and values using three separate prompts: one with the base model W_0 and combined prompt x, and one each with the content and style LoRA and their respective prompts (See Fig. 2). The resulting feature maps are then added together as

$$K \text{ or } V(x, x_s, x_c) = W_0^T x + \Delta W_s^T x_s + \Delta W_c^T x_c, \tag{3}$$

where x is the embedding for the combined prompt "A $<c>$ in $<s>$ style", while x_c and x_s are the embeddings of subject and style descriptors $<c>$ and $<s>$ respectively. This allows each LoRA to attend to different concepts. As illustrated in Fig. 5 (row M1), adding prompt separation to a DreamBooth baseline prevents cross-contamination and successfully destylizes the content.

3.3.2. Column Separation

Apart from learning separate concepts, we want our resulting LoRAs to be compatible, *i.e.* we want to be able to combine them through direct arithmetic merge to generate the subject and style together. Prompt separation effectively guides the two LoRAs to learn distinct concepts, and especially helps in achieving better destylization of the content. However, it does not guarantee compatibility since training LoRAs with different prompts can result in weight misalignment when they are combined to process the same prompt during inference.

To address this, we introduce the concept of column masks for each LoRA, denoted as m_s and m_c. These column masks dynamically control the contribution of each

column in the learned LoRA weights (see Fig. 2). Essentially, they allow the model to selectively activate or suppress specific columns within each LoRA, promoting orthogonality and reducing interference. By incorporating these column masks, the attention block update is modified as follows:

$$K \text{ or } V(x, x_s, x_c) = W_0^T x + m_s \Delta W_s^T x_s + m_c \Delta W_c^T x_c. \tag{4}$$

Sparse masks with importance re-calibration. To further decompose the concepts, instead of training the entire LoRA matrices, we find that training only a fraction of total columns in each weight matrix is sufficient to learn the concepts, and it ensures weight sparsity for improved decomposition. This strategy is inspired by Liu et al. [22], who find that a small set of neurons tend to be much more salient than the others for capturing concepts.

We use a dynamic approach to select the most important $N\%$ of the columns during training. Before the training, we initialize the column masks, m_s and m_c, with the top $(N/3)\%$ of the most important columns for style and content. Importance is calculated using the Cone method [22], with five warm-up training steps using the full LoRA weights. Then, during the remainder of training, for every t steps, we re-calibrate the column masks by calculating the column importance of LoRA weights and adding the new top $(N/3)\%$ of the most important columns to m_s and m_c until the $N\%$ cap is reached. In practice, N in the range of 25 to 40 work well, and we choose $N = 30$ for our experiments.

Orthogonal loss. To promote compatibility between the subject and style LoRAs, we leverage the following orthogonality loss on m_c and m_s:

$$\mathcal{L}_\perp = \sum_i |m_c^i \cdot m_s^i|, \tag{5}$$

minimizing which promotes orthogonality between the learned content and style weights [32]. Orthogonal loss $\mathcal{L}_\perp$ is added to our reconstruction loss $\mathcal{L}_{DB}$ as regularizer with the weight parameter $\lambda_\perp$.

This strategy brings significant improvements in compatibility of the resulting LoRAs, preventing overfitting to the input image and producing better recontextualizations, as shown in Fig. 5 (row M2).

3.3.3. Block Separation

B-LoRA [6] showed that certain blocks in the SDXL U-Net are more responsible for content and some are more responsible for style. We can leverage this insight by relaxing the column sparsity constraints on the style and subject LoRAs corresponding to the style-sensitive and subject-sensitive blocks, respectively. In other words, all the LoRA columns in these blocks are fully trained without sparse masks. In our attempt to take advantage of B-LoRA's block constraints, we find their split between subject and style is coarse-grained: tellingly, they fail to extract and represent the subject from the input image in its unstylized appearance. Therefore, we extend their approach by involving more blocks, and performing more fine-grained block-wise allocations of subject and style within the Up-blocks of SDXL U-Net (details in supplementary). This further improves the accuracy of fine details, especially for the style LoRA (see last row of Fig. 5).

4. Experiments

4.1. Implementation Details

Dataset. In our experiments, we use a set of 40 diverse images with unique styles and subjects. These are collected from previous work, state-of-the-art text-to-image generators, and open-source repositories. Attribution information is provided in the supplementary material.

Experimental setup. We use SDXL v1.0 [26] as our base model. Subject and style LoRAs are trained with rank $=$ 64 using Adam (learning rate = $5e-5$) for 600 steps with batch size $=$ 1, keeping the base SDXL weights and text encoders frozen. Column separation uses $t = 200$, $N = 30\%$, and weight of orthogonal loss is set to $\lambda_\perp = 0.5$ in all our experiments. Our block separation uses *all* upsampling blocks in the SDXL U-Net, unlike B-LoRA that uses just two. See supplementary material for details of which blocks we assign to content and which ones to style learning.

4.2. Qualitative Results

As shown in Fig. 1, the subject and style LoRAs obtained by UnZipLoRA can be used separately to generate new representations of subject-only and style-only concepts. A key advantage of UnZipLoRA is its ability to produce compatible subject and style LoRAs that can be seamlessly merged via direct addition. This allows for the generation of novel recontextualizations that faithfully incorporate *both* the subject and style of the original image. Fig. 3 demonstrates this capability through recontextualizations using either individual LoRAs or a combination of both. More results and comparisons are included in the supplementary material.

4.3. Comparative Evaluation

In this section, we compare UnZipLoRA with three recent methods: DreamBooth-LoRA [29], Inspiration Tree [37], and B-LoRA [6]. As shown in Fig. 4, our results are clearly superior qualitatively, and as a consequence, they are strongly preferred in a user study (Tab. 1).

For DreamBooth-LoRA [29], we train subject and style LoRA models separately using the DreamBooth method with the trigger phrases "A $<c>$" and "A $<c>$ in $<s>$ style." Note that the $<c>$ and $<s>$ used here are identical to those in UnZipLoRA. As can be seen from Fig. 4, al-

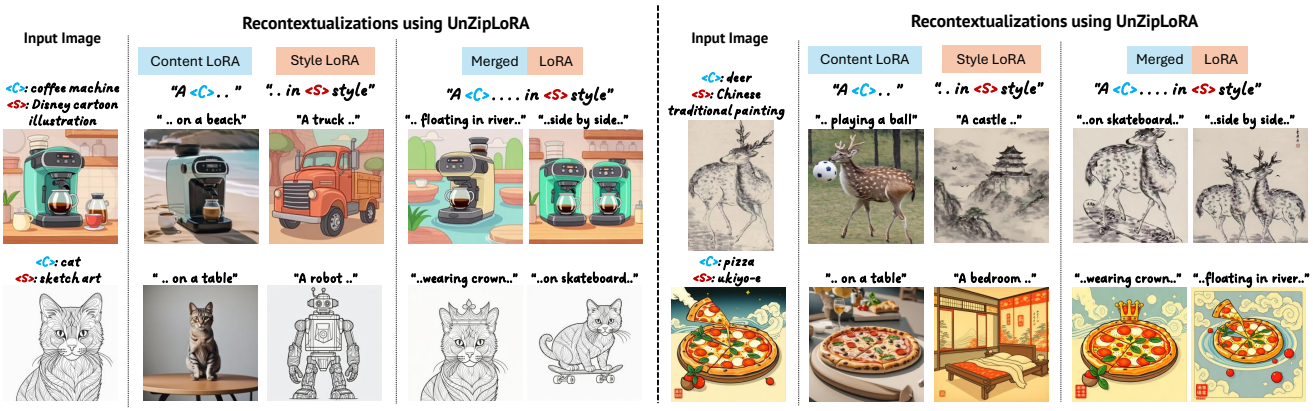

Figure 3. **Recontextualization.** The trained style and content LoRAs can be used individually or jointly at the inference time. The learned concepts can be used to generate images in various contexts, validating our method's robustness. Additional examples in the supplementary.

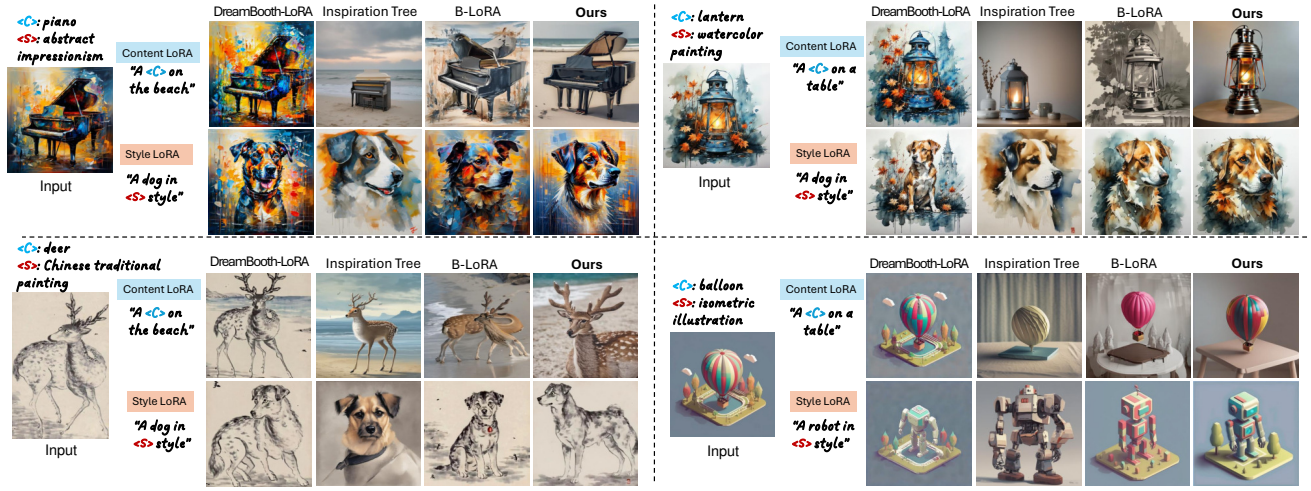

Figure 4. **Qualitative Comparisons.** Subject and style decomposition outputs from UnZipLoRA compared against DreamBooth-LoRA [29], Inspiration Tree [37], and B-LoRA [6]. UnZipLoRA achieves superior subject and style fidelity, disentangling and preserving these concepts more effectively. More examples can be found in the supplementary.

though DreamBooth-LoRA demonstrates satisfactory performance in learning the style, it fails to effectively destylize the subject and results in overfitting. This highlights the inherent difficulty in disentangling subject and style from a single image, even when training the LoRAs independently.

Inspiration Tree [37] is a concept extraction method that decomposes the input into a binary concept tree using Textual Inversion [8]. We train their model to extract subject and style using the prompt "A $<c>$ in $<s>$ style", and follow their default setup that initializes the placeholders $<c>$ and $<s>$ with "object" and "art" respectively. Despite employing separate training for content and style, Inspiration Tree struggles to disentangle them accurately. While it can identify the overarching category of concepts correctly, it fails to capture the intricate details of the subject or the style. This limitation stems from its reliance on Textual Inversion, which focuses on learning text embeddings rather than fine-tuning the weights, resulting in limited expressiveness and controllability.

We also compare our approach with B-LoRA [6] – a technique aimed at combining style and subject of two different images. We follow the default setup of B-LoRA, and independently train both the subject and style representations on the same input image using same trigger phrases as UnZipLoRA. We keep other hyperparameters (learning rate, training steps, etc.) consistent with [6]. As evident in the lantern and balloon examples in Fig. 4, B-LoRA outputs often retain residual style features and extraneous background elements, indicating overfitting and incomplete disentanglement. Furthermore, B-LoRA fails to accurately capture the color of the balloon, treating it as part of the style rather than the content. The limitations of B-LoRA

Table 1. **User Preference Study**. We compare the user preference for subject-style decomposition (top), and combined subject-style recontextualization (bottom) between our approach and competing methods. Users generally prefer our approach in both. We received 204 responses for each study from 34 total participants.

% **Preference for UnZipLoRA over:**			
	DreamBooth LoRA	Inspiration Tree	B-LoRA
Decomposition	91.17%	81.53%	62.74%
Recontextualization	98.10%	79.17%	77.14%

Table 2. **Subject-alignment and Style-alignment Scores.** Comparisons for Content and Style Decomposition.

	DB-LoRA	Insp. Tree	B-LoRA	UnZipLoRA
Style-align. (CLIP-I) ↑	0.417	0.404	0.418	**0.427**
Subject-align. (DINO) ↑	0.339	0.291	0.337	**0.349**
Style-align. (CSD) ↑	0.245	0.229	0.244	**0.265**
Subject-align. (CSD) ↑	0.338	0.334	0.342	**0.358**

can be attributed to its block-wise training approach, where content information is confined to specific blocks of the U-Net, potentially leading to information loss. By contrast, UnZipLoRA demonstrates more stable content destylization and better consistency in generating variations. Moreover, unlike B-LoRA, our method trains both the LoRAs jointly, reducing the compute requirements significantly.

Tab. 1 presents results of user studies comparing our method with the competing approaches for both subject/style decomposition and recontextualization. In our study, each participant is shown the input image, along with the outputs of two methods being compared (the methods are not labeled and their order is arbitrary). Each output group consists of 4 images for the subject and 4 images for style selected randomly, and participants are asked to choose an output group that decomposes the input image into style and content more accurately (see supplementary material for an example screenshot of the interface). We conducted three separate user studies – one for each competing method vs. UnZipLoRA – and received 204 responses for each from 34 total participants. As can be seen in Tab. 1, UnZipLoRA is strongly preferred over all three competing methods for both decomposition and recontextualization.

Tab. 2 further presents comparisons using automatic subject-alignment and style-alignment scores. We calculate alignment scores between decomposition output and the input image using standard CLIP-I image embeddings [27] for style and DINO features [2] for content. In both cases, we use cosine similarity as the metric and calculate averages over 8 input images by generating 16 samples for each. As can be seen from Tab. 2, UnZipLoRA achieves the highest alignment scores for both subject and style, highlighting its superiority. DreamBooth-LoRA tends to overfit the input image, resulting in higher scores even though its outputs are qualitatively inferior to those of other methods.

CLIP-I and DINO metrics are inherently limited, especially in measuring style alignment, since they may not fully capture stylistic nuances and can be influenced by the semantic content of the images. As a more sensitive metric, we use a recently proposed CSD model [34] trained specifically to extract the style descriptors from images. We use the embeddings from CSD's content and style branches to measure subject and style fidelity respectively. The CSD alignment scores (third and fourth line of Tab. 2) yield a clearer separation between our method and the others.

4.4. Ablation Study

In this section, we present an ablation study to justify the effectiveness of our system design. As illustrated in Fig. 5, we start from DreamBooth [29] as the baseline and successively add the key components of our method: prompt separation (M1), column separation (M2), and block separation (M3). Tab. 3 reports user preferences for each model version over the previous one (based on 204 responses from 34 participants).

As can be seen from Fig. 5, the DreamBooth baseline is incapable of subject-style disenganglement and has very weak recontextualization ability. Adding **prompt separation** (**M1**) to the baseline enables successful extraction of a realistic content LoRA and improves recontextualization. However, prompt separation alone struggles to capture all the complexities of the style from the input image (only 12.35% user preference over baseline for style decomposition), and shows strong overfitting to the input image in combined recontextualization, indicating incompatibility of the learned LoRAs. Adding **column separation** (**M2**) on top of **M1** reduces the interference between subject and style LoRAs while also improving style generation abilities. For example, details like the color of the microscope in Fig. 5 are retained, and the combined recontextualization performance improves significantly. However, improvements in style decomposition remain modest, indicating that using a small portion of the columns for style is insufficient for effective style learning. Finally, adding **block separation** (**M3**) with subject- and style-specific blocks on top of **M2** significantly enhances the model's ability to capture fine-grained stylistic features. The last row of Fig. 5 shows that using all three separation strategies together achieves the best results, enabling comprehensive disentanglement of subject and style and flexible recontextualization.

4.5. Applications and extensions

Cross-combination of subject and style LoRAs from different images. The subject and style LoRAs produced by UnZipLoRA open up a possibility for cross-combination: pairing a subject LoRA from one image with a style LoRA from another. Fig. 6 shows such cross-combination results where the LoRAs are combined by direct addition. While

Table 3. **Ablation User Study**. **Baseline**: Dreambooth-LoRA; **M1**: Prompt-separation only; **M2**: Prompt- and column- separations; **M3 (full)**: Prompt-, column-, and block- separations.

	% **Preference for:**		
	M1 over Baseline	**M2 over M1**	**M3 over M2**
Subject Decomposition	91.67%	55.74%	**55.36%**
Style Decomposition	12.35%	39.51%	**86.42%**
Combined Recontextualization	92.80%	93.64%	**61.90%**

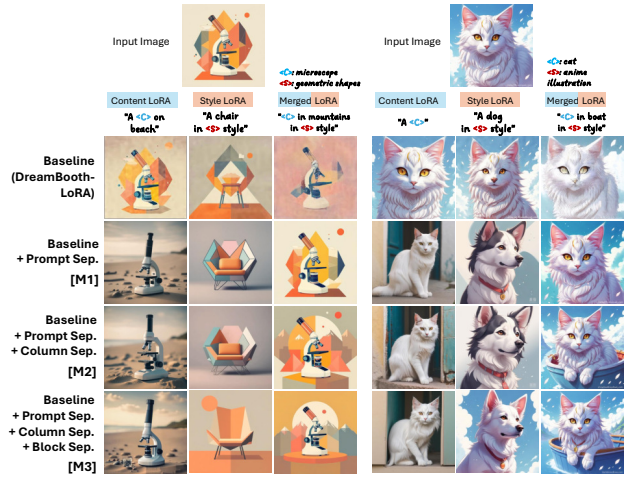

Figure 5. **Illustration of ablations.** Each column shows the performance of subject LoRA, style LoRA, and combined LoRA.

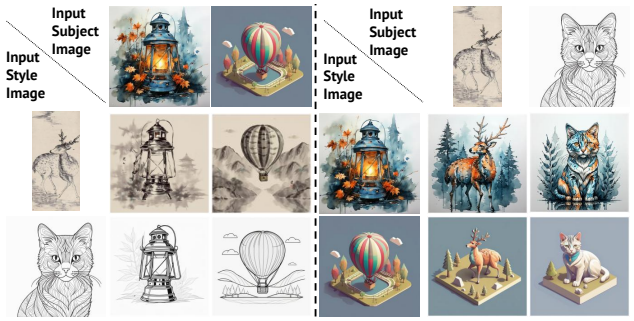

Figure 6. UnZipLoRA provides the valuable capability of *cross-composition* using subject and style LoRAs from different images.

these LoRAs are not explicitly trained together (and thus not subject to the orthogonality constraints enforced by ZipLoRA [32]), the inherent separation imposed by our column and block strategies generally results in higher compatibility than generic DreamBooth-LoRAs trained without such constraints. Consequently, direct arithmetic merger yields promising cross-stylization results.

Extension to other architectures. To demonstrate the possibility of extending our method beyond SDXL, we train on KOALA [19], a more efficient recent model with a leaner U-Net architecture. As shown in Fig. 7, our method, when applied to KOALA, accurately captures subject and style and allows for successful recontextualization (though the overall quality of the results is not as high as for SDXL due to limited capacity and lower parameter count of KOALA). The core idea of UnZipLoRA, that of simultaneously training two LoRAs on the same input image, should also be applicable to other architectures like DiT [25], though this extension is beyond the scope of present work.

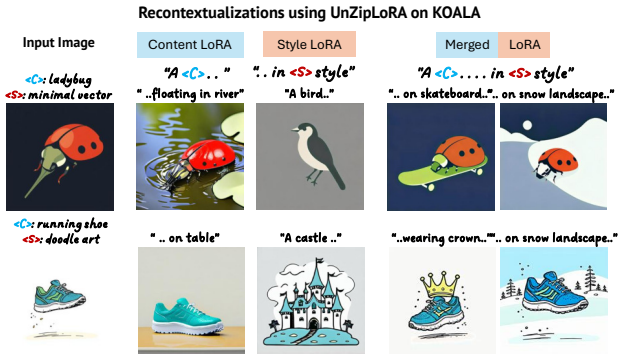

Figure 7. Our approach generalizes effectively, as demonstrated by successful results on the more recent KOALA diffusion model.

Figure 8. **Failure Cases.** In a few cases with highly abstract styles, UnZipLoRA may fail to destylize the subject accurately (note the unnatural shape of the bear). B-LoRA exhibits a similar failure.

5. Conclusion

This paper introduced UnZipLoRA, a novel method for decomposing a single image into disentangled, compatible subject and style LoRAs. UnZipLoRA utilizes prompt, column, and block separation strategies to effectively extract these elements, enabling diverse recontextualizations and manipulations. Our experiments demonstrate superior performance compared to existing methods, highlighting UnZipLoRA's potential for creative exploration and control within text-to-image generation. While robust across a wide range of subjects and styles, UnZipLoRA may fail to destylize the subject when the style involves too much shape distortion or abstraction, as observed in Fig. 8. Future work includes exploring alternative disentanglement techniques for such challenging cases, training-free approaches for improved efficiency, and extending UnZipLoRA to other architectures for better generalization.

Acknowledgments. This research was supported in part by NSF grant CCF 2348624.

References

[1] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. *SIGGRAPH Asia 2023 Conference Papers*, 2023. 3

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 7

[3] Min Jin Chong and David A. Forsyth. Jojogan: One shot face stylization. *CoRR*, abs/2112.11641, 2021. 3

[4] Ziyi Dong, Pengxu Wei, and Liang Lin. Dreamartist: Towards controllable one-shot text-to-image generation via positive-negative prompt-tuning, 2023. 3

[5] Alexei A Efros and William T Freeman. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 341–346, 2001. 3

[6] Yarden Frenkel, Yael Vinker, Ariel Shamir, and Daniel Cohen-Or. Implicit style-content separation using b-lora. In *ECCV*, 2024. 2, 3, 5, 6

[7] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ArXiv*, abs/2108.00946, 2021. 3

[8] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022. 2, 3, 6

[9] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 3

[10] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305*, 2023. 2

[11] Shaozhe Hao, Kai Han, Zhengyao Lv, Shihao Zhao, and Kwan-Yee K. Wong. ConceptExpress: Harnessing diffusion models for single-image unsupervised concept extraction. In *ECCV*, 2024. 3

[12] Aaron Hertzmann, Charles E. Jacobs, Nuria Oliver, Brian Curless, and D. Salesin. Image analogies. *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001. 3

[13] Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239, 2020. 2, 3

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 2, 3

[15] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 3

[16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016. 3

[17] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 2, 3

[18] Gihyun Kwon and Jong-Chul Ye. One-shot adaptation of gan in just one clip. *ArXiv*, abs/2203.09301, 2022. 3

[19] Youngwan Lee, Kwanyong Park, Yoorhim Cho, Yong-Ju Lee, and Sung Ju Hwang. Koala: Empirical lessons toward memory-efficient and fast diffusion models for text-to-image synthesis, 2023. 2, 3, 8

[20] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *Advances in Neural Information Processing Systems*, 2017. 3

[21] Mingcong Liu, Qiang Li, Zekui Qin, Guoxin Zhang, Pengfei Wan, and Wen Zheng. Blendgan: Implicitly gan blending for arbitrary stylized face generation. *ArXiv*, abs/2110.11728, 2021. 3

[22] Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized generation. *arXiv preprint arXiv:2303.05125*, 2023. 5

[23] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A. Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10738–10747, 2021. 3

[24] Dae Young Park and Kwang Hee Lee. Arbitrary style transfer with style-attentional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5880–5888, 2019. 3

[25] William Peebles and Saining Xie. Scalable diffusion models with transformers. *ICCV*, 2023. 8

[26] Dustin Podell, Zion English, Kyle Lacey, A. Blattmann, Tim Dockhorn, Jonas Muller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *ArXiv*, abs/2307.01952, 2023. 3, 5

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7

[28] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021. 2, 3

[29] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 2, 3, 5, 6, 7

[30] Simo Ryu. Merging loras. https://github.com/cloneofsimo/lora. 3

[31] Viraj Shah, Ayush Sarkar, Sudharsan Krishnakumar Anita, and Svetlana Lazebnik. Multistylegan: Multiple one-shot image stylizations using a single gan. *arXiv*, 2023. 3

[32] Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. ZipLoRA: Any subject in any style by effectively merging LoRAs. In *ECCV*, 2024. 2, 3, 5, 8

[33] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*, 2023. 2, 3

[34] Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models. *ArXiv*, abs/2404.01292, 2024. 7

[35] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ArXiv*, abs/2010.02502, 2020. 2, 3

[36] Joshua Tenenbaum and William Freeman. Separating style and content. In *Advances in Neural Information Processing Systems*, 1996. 1

[37] Yael Vinker, Andrey Voynov, Daniel Cohen-Or, and Ariel Shamir. Concept decomposition for visual exploration and inspiration. *ACM Transactions on Graphics (TOG)*, 42:1 – 13, 2023. 2, 3, 5, 6

[38] Yue Wang, Ran Yi, Ying Tai, Chengjie Wang, and Lizhuang Ma. Ctlgan: Few-shot artistic portraits generation with contrastive transfer learning. *ArXiv*, abs/2203.08612, 2022. 3

[39] You Wu, Kean Liu, Xiaoyue Mi, Fan Tang, Juan Cao, and Jintao Li. U-vap: User-specified visual appearance personalization via decoupled self augmentation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9482–9491, 2024. 3

[40] Zhi Xu, Shaozhe Hao, and Kai Han. Cusconcept: Customized visual concept decomposition with diffusion models. *ArXiv*, abs/2410.00398, 2024. 2, 3

[41] Ceyuan Yang, Yujun Shen, Zhiyi Zhang, Yinghao Xu, Jiapeng Zhu, Zhirong Wu, and Bolei Zhou. One-shot generative domain adaptation, 2021. 3

[42] Shuai Yang, Liming Jiang, Ziwei Liu, and Chen Change Loy. Pastiche master: Exemplar-based high-resolution portrait style transfer, 2022.

[43] Peihao Zhu, Rameen Abdal, John Femiani, and Peter Wonka. Mind the gap: Domain gap control for single shot domain adaptation for generative adversarial networks. In *International Conference on Learning Representations*, 2022. 3