Diffusion Image Analogies

Adéla Šubrtová
CTU in Prague, FEE
Prague, Czech Republic
subrtade@fel.cvut.cz

Michal Lukáč
Adobe Research
San Jose, California, United States
lukac@adobe.com

Jan Čech
CTU in Prague, FEE
Prague, Czech Republic
cechj@fel.cvut.cz

David Futschik
CTU in Prague, FEE
Prague, Czech Republic
futscdav@fel.cvut.cz

Eli Shechtman
Adobe Research
Seattle, Washington, United States
elishe@adobe.com

Daniel Sýkora
CTU in Prague, FEE
Prague, Czech Republic
sykorad@fel.cvut.cz

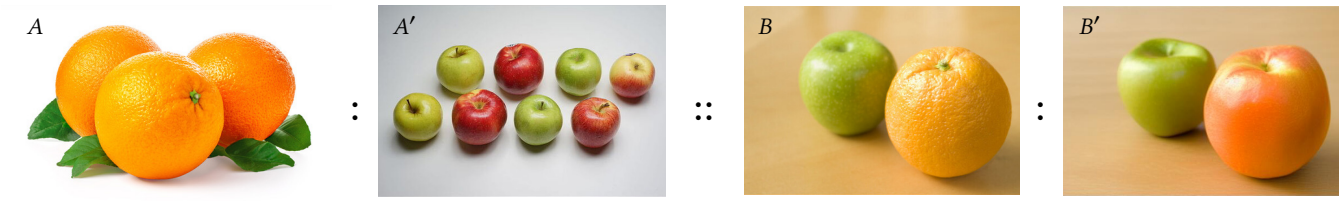

Figure 1: Diffusion Image Analogies in action—a pair of input images A and A' defines an analogy, i.e., a semantic transition according to which the target image B is modified to produce the output image B'. In this case the user's intention is to change oranges to apples. Note, how our method captures the analogy implicitly without the need to have objects aligned in a similar scale or provide an additional guidance which is a vital requirement for previous techniques based on image analogies paradigm [Hertzmann et al. 2001]. Source images: Adobe Stock A, © Dllu A', © The Busy Brain B.

ABSTRACT

In this paper we present Diffusion Image Analogies—an example-based image editing approach that builds upon the concept of image analogies originally introduced by Hertzmann et al. [2001]. Given a pair of images that specify the intent of a specific transition, our approach enables to modify the target image in a way that it follows the analogy specified by this exemplar. In contrast to previous techniques which were able to capture analogies mostly on the low-level textural details our approach handles also changes in higher level semantics including transition of object domain, change of facial expression, or stylization. Although similar modifications can be achieved using diffusion models guided by text prompts [Rombach et al. 2022] our approach can operate solely in the domain of images without the need to specify the user's intent using textual form. We demonstrate power of our approach in various challenging scenarios where the specified analogy would be difficult to transfer using previous techniques.

CCS CONCEPTS

• **Computing methodologies** → **Image processing**.

SIGGRAPH '23 Conference Proceedings, August 06–10, 2023, Los Angeles, CA, USA

ACM ISBN 979-8-4007-0159-7/23/08.
https://doi.org/10.1145/3588432.3591558

KEYWORDS

image analogies, diffusion networks, large language models

ACM Reference Format:
Adéla Šubrtová, Michal Lukáč, Jan Čech, David Futschik, Eli Shechtman, and Daniel Sýkora. 2023. Diffusion Image Analogies. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Proceedings (SIGGRAPH '23 Conference Proceedings), August 06–10, 2023, Los Angeles, CA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3588432.3591558

1 INTRODUCTION

In 2001 Hertzmann et al. [2001] pioneered the idea of image analogies where a given pair of images (denoted A and A') defines a visual relationship, which can be applied to a query image B to create an analogy B'. While Hertzmann et al. used a simple pixel window-based approach to realize the analogy, in the following decades this concept became quite popular. It inspired an entire family of methods which developed more sophisticated algorithms for more general approaches of guided image synthesis [Barnes et al. 2009; Bénard et al. 2013; Diamanti et al. 2015; Fišer et al. 2016; Freeman et al. 2002; Futschik et al. 2021; Jamriška et al. 2019; Ritter et al. 2006; Texler et al. 2020b] which became a powerful tools for various practical scenarios (image/video editing, stylization, or completion). A key limitation of all these methods is that they transfer appearance at quite low (pixel) level and have little notion of higher level context or ability to modify higher-level structure. High-level contextual analogies, such as presenting a no-smile/smile pair on the input and make a person in the target photo smile is difficult to achieve.

Recently, rapid development of large language models [Brown et al. 2020] paired with images [Radford et al. 2021] and combined with the power of diffusion models [Ho et al. 2022] lead to a revolutionary text-guided image generation approach [Nichol et al. 2022]. Systems such as Stable Diffusion [Rombach et al. 2022] demonstrate a breathtaking ability to govern high-level semantics and generate photorealistic as well as artistic images of high quality, following the users' specification. Upon this success text-based guidance gained significant popularity and prompt engineering become a new content creation skill. Although in less restrictive scenarios text prompts provide sufficient power to specify content that needs to be generated when going more into specific details the process may become difficult and leads to tedious trial-and-error workflow. Even longer text prompts may not sufficiently describe the user's intent so that the adage "a picture is worth a thousand words" become highly relevant.

In our work we aim to overcome above-mentioned difficulties by elevating the original concept of image analogies into a next level where also high-level contextual information is taken into account. To achieve this goal we employ the power of diffusion models and demonstrate how to achieve image-based analogy without the need to rely on text prompts yet still being able to leverage the ability of large language models to govern high-level semantics.

A key idea behind our method is to invert guiding/content images to get their initial random noise and conditioning matrix which can then be used to closely reproduce the original images using diffusion process. Then we leverage a principle known from natural language processing [Mikolov et al. 2013] where algebraic operations applied to vector representations of input words can lead to equations such as: $\text{King} + (\text{Woman} - \text{Man}) = \text{Queen}$. We demonstrate that a similar approach can be used in the context of diffusion models where the conditioning matrix (originally extracted from a given text prompt) serves as a vector representing the semantics of the input image and thus can be used to express and apply high-level analogy.

2 RELATED WORK

The concept of *image analogies* [Hertzmann et al. 2001] was introduced in an application setting where an exemplar pair of images (the unfiltered and filtered) is used to synthesize a filtered version of a given target image. A guided texture synthesis algorithm similar to that proposed by Ashikhmin [2001] is used to perform the transfer that enables to produce analogies only with respect to low level details. Moreover, since only color information is used for guidance the output can suffer from ambiguity. One of the extensions Hertzmann et al. proposed to mitigate this issue was the texture-by-numbers concept where instead of a regular image a customly created guiding channel (e.g., a segmentation map) is used to define the analogy. This technique inspired others to develop more advanced guided patch-based synthesis [Diamanti et al. 2015; Fišer et al. 2016, 2017; Zhou et al. 2017] that can deliver compelling results in more complex scenarios where additional contextual information is necessary, e.g., when synthesizing an image that respects prescribed depth, illumination, weathering effects or facial expressions. Nonetheless, guiding channels that inject such semantics need to be prepared explicitly in advance and are application specific.

Gatys et al. [2016] proposed a complementary approach to image analogies called *neural style transfer*. In this technique instead of specifying a full analogy only the style image is provided. The assumption here is that responses of VGG network trained on object recognition tasks [Simonyan and Zisserman 2014] can distill higher level semantics as well as low-level details so that one can use them to guide the synthesis and combine visual characteristics of the input style with the content of the target scene. This approach was later extended by others [Kolkin et al. 2019; Li et al. 2017] and combined with the principles of patch-based synthesis [Li and Wand 2016; Liao et al. 2017; Texler et al. 2020a] that use responses of VGG explicitly as a latent guiding channel. Although these techniques can deliver impressive results in various practical scenarios their ability to capture higher level context is bounded by capabilities of VGG network and do not allow for more generic customization that is possible in the concept of image analogies.

Above mentioned limitations can be to some extent addressed by *image-to-image translation* methods [Futschik et al. 2019; Isola et al. 2017; Park et al. 2020; Zhu et al. 2017] where the user specifies a larger dataset of exemplar pairs as an input. Those are then used to train a translation network that can deliver expected modification of the target image. A key drawback here is that the creation of such large input dataset can be non-trivial and also the training phase requires huge computational overhead that can be prohibitive in the case where the translation domain is not known in advance.

To overcome the necessity of larger set of aligned image pairs methods have been proposed to handle unpaired exemplars [Zhang et al. 2020; Zhu et al. 2017] as well as notably smaller datasets [Liu et al. 2019] that in specific cases can be reduced to a single pair of images [Futschik et al. 2021]. Nonetheless, since those techniques do not consider larger contextual model to capture higher-level semantics they still either transfer only lower level textural details or require a few more examples to distill the analogy statistically.

Thanks to the advent of CLIP [Radford et al. 2021] that can employ large language model [Brown et al. 2020] to measure how closely a given text describes an input image, *text-guided image generation* [Gal et al. 2022b; Patashnik et al. 2021] started to gain significant popularity namely in combination with diffusion models [Ho et al. 2022]. Systems such as Stable Diffusion [Rombach et al. 2022] become a revolutionary new tool for guided image synthesis and editing [Avrahami et al. 2022]. To enable local editing Hertz et al. [2022] inject attention maps from initially generated image to retain the original structure. To preform the edit on real images diffusion process needs to be inverted using DDIM [Song et al. 2021]. Mokady et al. [2022] improve on DDIM using null-text optimization where only the unconditional textual embedding is modified. Kawar et al. [2023] instead fine-tune the diffusion model to capture appearance of the input image. Finally, Brooks et al. [2023] bypass the inversion by training a conditional diffusion model using large synthetically generated dataset of image editing examples.

Despite the unprecedented quality of results produced by techniques mentioned above their key limitation is that their guidance is dependent on specifying text prompts. Our aim is to relax such a requirement and get closer to the idea of *visual prompting* [Bar et al. 2022] where instead of text, analogy expressed by two images is provided to guide the transfer yet contrary to the work of Bar et al. in our approach we retain the semantic power of CLIP. A similar

approach appeared recently in the works of Tumanyan et al. [2022] and Kwon and Ye [2023] who let the user specify an appearance exemplar and use another image to define the target structure. Semantic transfer is then performed using deep features of pre-trained vision transformer [Caron et al. 2021]. Ruiz et al. [2023] instead fine-tune a diffusion model by binding a unique identifier to a specific subject represented by a set of input images. Despite impressive results those techniques produce they do not address image analogies scenario.

Our approach bears resemblance to the work of Tewel et al. [2022] who infer a caption for a given input image by combining visual-semantic model (CLIP) with large language model (GPT-2) [Radford et al. 2019]. Tewel et al. demonstrate that hybrid image/text analogy puzzles can be solved by using simple arithmetic operations between estimated CLIP features; however, the output is always a text and not an image.

3 OUR APPROACH

Similarly to Hertzmann et al. [2001] the input to our method is a triplet of images A, A', and B where A to A' represent the intended analogy and the aim is to modify the image B to produce an image B' in such a way that the change performed follows the analogy represented by A and A', i.e., formally: $A : A' :: B : B'$ (see Fig. 1). In the solution proposed by Hertzmann et al. objects and other structures in images A and A' are assumed to be spatially aligned and the change between them mostly happens on a pixel level. In our case we consider arbitrary images that may not be spatially aligned and focus more on higher level context. This shifts the problem setting closer to deep image analogies [Liao et al. 2017] or a generic style transfer scenario [Gatys et al. 2016], however, in contrast to those previous approaches our solution provides a more explicit control over the transfer by providing the analogy $A : A'$. Moreover, since our aim is to depart from pixel level features towards higher level contextual information we employ CLIP [Radford et al. 2021] that can interconnect the image with semantically meaningful prior of large language model.

To achieve this goal we initially assume the input images A, A', and B were generated synthetically using Stable Diffusion [Rombach et al. 2022]. Here a diffusion process $\mathbf{U}^+$ is used to produce images that follows a given text descprition. $\mathbf{U}^+$ consists of multiple denoising steps where a pre-trained U-net $\mathbf{U}$ is applied repeatedly on an initial noise image ϵ_* while CLIP features c_* derived from the input text prompt are used to guide the diffusion, i.e.: $A = \mathbf{D}(\mathbf{U}^+(\epsilon_A, c_A))$, $A' = \mathbf{D}(\mathbf{U}^+(\epsilon_{A'}, c_{A'}))$, and $B = \mathbf{D}(\mathbf{U}^+(\epsilon_B, c_B))$. Since Stable Diffusion operates in latent representation with reduced dimensionality we need to use decoder $\mathbf{D}$ proposed by Rombach et al. to reconstruct the final image.

Similarly to Tewel et al. [2022] or Ramesh et al. [2022] we express the analogy $e \approx A : A'$ algebraically by subtracting CLIP features of images A and A', i.e., $e_{A:A'} = c_{A'} - c_A$. The output image B' can then be produced using diffusion process $\mathbf{U}^+$ and decoder $\mathbf{D}$:

$$B' = \mathbf{D}(\mathbf{U}^+(\epsilon_B, c_B + \lambda \cdot e_{A:A'})), \tag{1}$$

where λ is a hyper-parameter that describes the strength of the analogy.

Such an operation can trivially be applied to images A, A', and B that were produced synthetically using Stable Diffusion. However, a key question here is how to apply the same process to real images for which the initial noise ϵ_* as well as CLIP features c_* are not known.

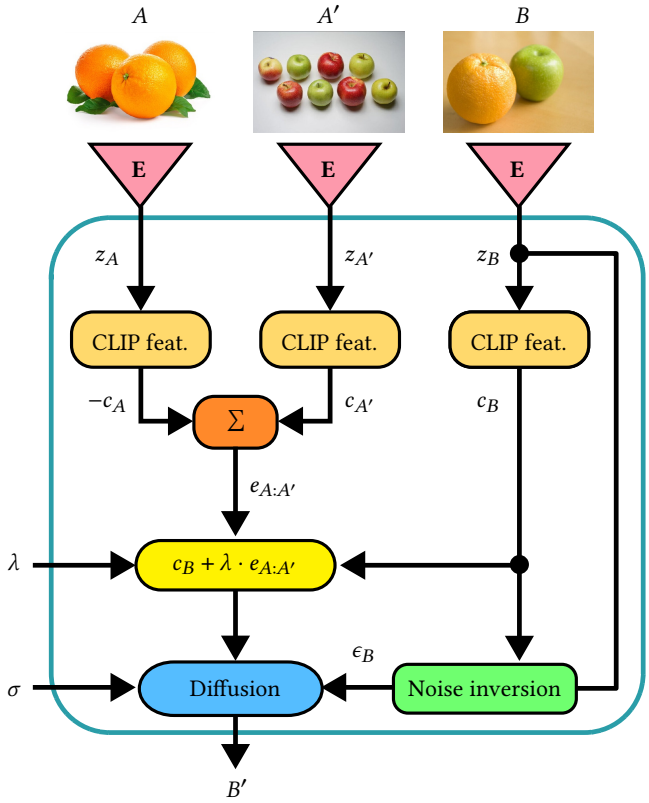

Figure 2: Diffusion Image Analogies pipeline—input images A, A', and B are fed into Stable Diffusion encoder E [Rombach et al. 2022] to produce their lower dimensional latent representations z_A, $z_{A'}$, and z_B. Subsequently, their corresponding CLIP features c_A, $c_{A'}$, and c_B are estimated (c.f. Section 3.1) and the analogy factor $e_{A:A'}$ is computed. Initial noise image ϵ_B is estimated (c.f. Section 3.2) from c_B and z_B and the scaled factor $\lambda \cdot e_{A:A'}$ is added to the CLIP features c_B to drive the diffusion process $\mathbf{U}^+$ that is initiated by ϵ_B. Once the diffusion process is completed in the latent space the final image B' is reconstructed from the latent space using Stable Diffusion decoder D (included in the Diffusion block). The user can influence the entire process by manipulating analogy strength parameter λ and guidance scale parameter σ. Source images: Adobe Stock A, © Dllu A', © The Busy Brain B.

In our solution (c.f. Fig. 2) we infer ϵ_* and c_* using optimization. A key challenge here is that both the initial noise image as well as CLIP features influence the image formation process jointly through the diffusion process $\mathbf{U}^+$, albeit not interchangeably. A naive joint optimization scheme would lead to an unstable solution, where the image is reconstructed well, but the semantic information is not cleanly factored out into c_*. In such a configuration, minor perturbations to ϵ_* cause the diffusion to produce broken or nonsensical images (see Fig. 8a).

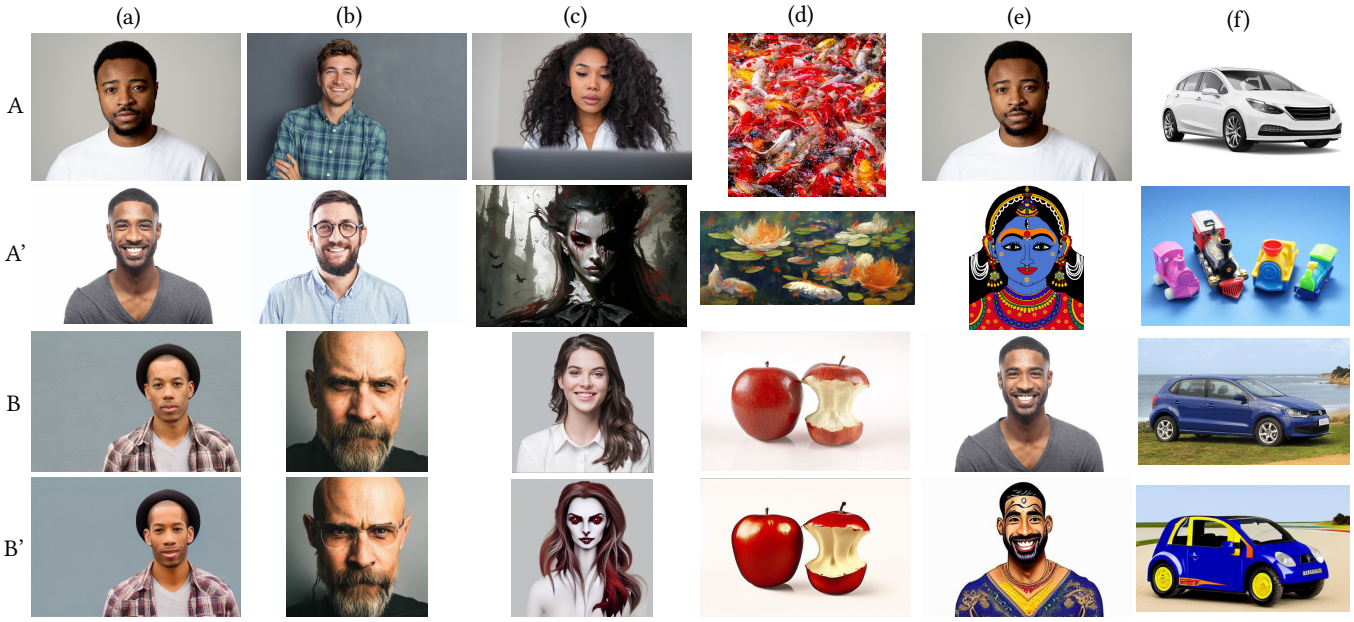

Figure 3: Results of Diffusion Image Analogies—the input analogy $A : A'$ is applied to the target image B producing the output image B'. The analogies are used to: (a) make a person smile, (b) make a person wear glasses, (c) make a woman to look like a vampire, (d) stylize an input photo, (e) make a man to look like an Indian god, (f) make a real car to look like a plastic toy. Source images: © Kevin Bidwell (b) B, © Lucíola Correia (d) B, © The NRMA (f) B, Adobe Stock the rest.

The reason for this is that information from the training data is not distributed uniformly over the space of all c_*, or even evenly within a given c_* matrix. Due to how CLIP is structured, information is mostly concentrated in the upper rows of c_*, but because unconstrained optimization is free to use the entire tensor, it can recover a result that reconstructs the image faithfully, but does not correspond to any meaningful sequence of text tokens. Such solutions necessarily lie in a subspace where $\mathbf{U}$ is not well trained, and therefore the output is not robust to ϵ.

To avoid such an instability, we need several layers of regularization. Firstly, we recover the CLIP features and the initial noise in a sequence rather than jointly. Then, we architect the optimization and apply a regularization described below to ensure the recovered features mimic meaningful word token sequences and actually capture most of the high level features, thereby making the output more robust to the initial noise.

3.1 Estimation of the CLIP Features

Although the CLIP feature parameters c_* can be estimated directly there is no guarantee they will contain desired semantic information about the input image (see Fig. 8b). To alleviate this drawback we regularize their values by plugging CLIP model $\mathbf{C}$ into the optimization process. To do that we do not estimate c_* directly instead we optimize a set of tokens $\mathcal{K}$ that serves as an input to CLIP model from which the desired features can be produced: $c_* = \mathbf{C}(\mathcal{K})$. A similar idea was recently proposed by Gal et al. [2022a], however, in their solution only a single token (used in a variety of different sentences) is optimized. We extend this estimation to handle multiple tokens at the same time without any prior context. This enables us to to distill a more information-rich embedding c_*.

To estimate c_* for an image I we aim to minimize the loss:

$$\min_{\mathcal{K}} \sum_{t \in \mathcal{T}} |\epsilon - \mathbf{U}(z_t, t, \mathbf{C}(\mathcal{K}))|_2^2 . \tag{2}$$

Here $\mathcal{T}$ denotes individual steps of the diffusion process. We construct a set of noisy images $z_t = \sqrt{\alpha_t}\mathbf{E}(I) + \sqrt{1-\alpha_t}\epsilon$, $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, according to the model's noise schedule, and a feature encoding $\mathbf{E}(I)$ of the image I, as per Rombach et al. [2022]. Then we optimize for a sequence $\mathcal{K}$ which, when encoded as $\mathbf{C}(\mathcal{K})$, conditions the network $\mathbf{U}$ to predict the correct noise sample.

An essential part of our loss is that we restrict the number of tokens $\mathcal{K}$ being optimized from the original length of 77 to a smaller number N. Remaining tokens are set to be end-of-text tokens. The aim of this restriction is to encourage brevity and prioritize tokens that represent more salient high-level information. In the ablation experiments we demonstrate that such a sharpening effect has positive impact on the quality of the resulting CLIP features (see Fig. 13).

To further regularize the estimation of c_*, we augment the set of noisy images with images that are slightly geometrically modified versions of the original image. We use horizontal flip, random scale, and translation (compare Figures 8c and 8d to see the positive effect of this augmentation).

3.2 Estimation of the Initial Noise Image

In order to apply the analogy to the given image B we need to estimate its initial Gaussian noise ϵ_B so that $B = \mathbf{D}(\mathbf{U}^+(\epsilon_B, c_B))$.

Figure 4: Results of Diffusion Image Analogies (cont.)—the input analogy $A : A'$ is applied to the target image B producing the output image B'. The analogies are used to: (a) make a bear to look like a panda, (b) transform three cats to three dogs, (c) make a group of people practicing yoga to look like a group of Buddha statues. Source images: © Artanisen (a) A, © Cliff (a) A', © George Hodan (b) B, Adobe Stock the rest.

Figure 5: Comparison with the original Image Analogies [Hertzmann et al. 2001]. The intention of the given analogy $A : A'$ is to perform colorization. Despite the fact images A and A' are aligned (a requirement of the original method), the output is far from the original intention. Here the reason is that the method of Hertzmann et al. use intensity-based matching to retrieve similar features which is not semantically meaningful in this case. Using our approach the target image B is colorized appropriately. Source images: © Subhamshome28 B, Adobe Stock the rest.

Thanks to already estimated CLIP features c_B we can optimize ϵ_B through the entire diffusion process $\mathbf{U}^+$ by minimizing the loss:

$$\min_{\epsilon_B} \left|z_B - \mathbf{U}^+(\epsilon_B, c_B)\right|_2^2 + \left|B - \mathbf{D}(\mathbf{U}^+(\epsilon_B, c_B))\right|_2^2. \quad (3)$$

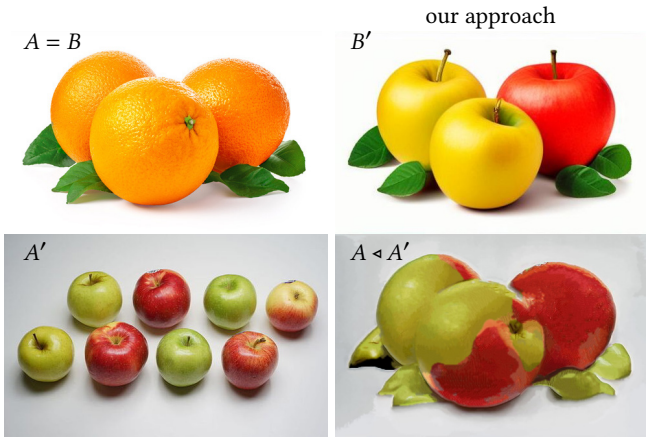

Figure 6: Comparison with Deep Image Analogies [Liao et al. 2017]. The method of Liao et al. does not support full image analogies scenario. Only two images $A : A'$ can be specified with the aim to transfer visual attributes between them $A \triangleleft A'$. To simulate full image analogies $A : A' :: B : B'$ we set $B = A$. From the resulting image B' it is visible that our approach handles high-level semantics better than the method of Liao et al. which tends to transfer only low level textural details. Source images: Adobe Stock A, © Dllu A'.

Here z_B is a latent representation of image B, i.e., $z_B = \mathbf{E}(B)$ and $\mathbf{U}^+$ is the entire diffusion process that starts with the initial noise $z_M = \epsilon_B$ at a time $i = M$. Then z_i is refined $z_{i-1} = \gamma_i z_i + \gamma_i' \mathbf{U}(z_i, i, c_B)$ until $i = 0$. Constants γ_i and γ_i' are scaling factors associated with $\mathbf{U}$. Note that we use the short schedule mode that relates to the original schedule $t \in (0, \ldots, T)$ as $t = i\frac{T}{M}$, skipping T/M denoising steps. Besides minimizing difference of latent representations we also minimize the difference of the original image B to its reconstructed and decoded counterpart $\mathbf{D}(\mathbf{U}^+(\epsilon_B, c_B))$.

4 RESULTS

We implemented our approach in Python using source code of Stable Diffusion [Rombach et al. 2022] as a basis (see our GitHub repository: https://github.com/subrtadel/DIA). For the estimation of CLIP embedding (2) we employed AdamW optimizer [Loshchilov and Hutter 2019] and we set the number of active tokens $N = 10$. On the GPU (Tesla A100 with 40 GB of RAM) this process takes approximately 8 minutes for 512×512 image. For the estimation of initial Gaussian noise ϵ_B of image B we use $M = 10$ steps of the diffusion process $\mathbf{U}^+$ ($T = 1000$) to minimize the loss (3) using L-BFGS solver [Berahas et al. 2016]. On the same GPU and image resolution this step takes around 9 minutes.

As soon as all necessary parameters (c_A, $c_{A'}$, c_B, and ϵ_B) are estimated we can compute the analogy (1). To enable better control over the final result we first calculate images B' for different values of the analogy strength λ (each sample takes around 0.7 secs on the GPU) and then we let the user to choose its optimal setting interactively. Another value the user can fine tune is the guidance

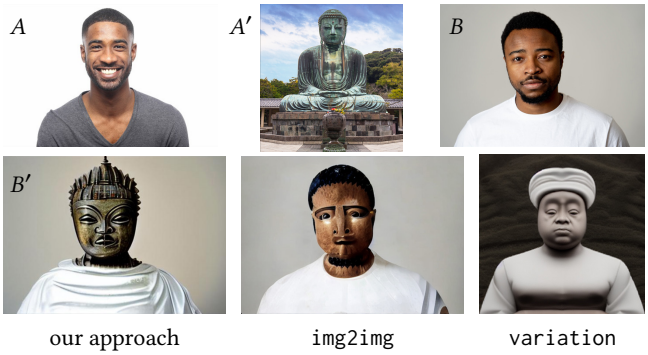

Figure 7: Comparison with Stable Diffusion [Rombach et al. 2022]. In our synthetic baseline BLIP [Li et al. 2022] is used to estimate text prompts of the images from which CLIP features are computed. In this case the output is: *"a man with a smile on his face"* for A, *"a big statue in a Chinese temple with an overhang"* for A', and *"a man in a white shirt looking at the camera"* for B. Those text prompts are used to produce CLIP features c_A, $c_{A'}$, and c_B from which the CLIP features of c'_B are computed: $c_{B'} = c_B + \lambda \cdot (c_{A'} - c_A)$. Then a `img2img` mode of Stable Diffusion is used and conditioned with $c_{B'}$ to produce the output B'. Note, how the analogy derived from the estimated text prompts does not capture semantics properly. In the `variations` mode a different model of Stable Diffusion is used that is conditioned on images directly. The analogy as well as the output conditioning with the image embedding is computed using the same approach as with text prompts. Note, that in this case the analogy is getting closer to the original intention, nevertheless, the overall structure is not preserved well. Source images: Adobe Stock.

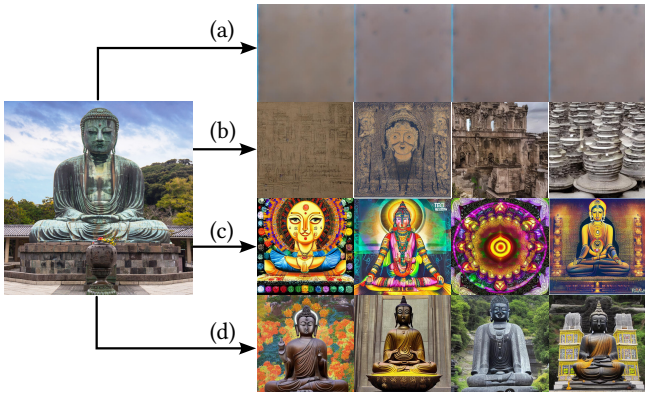

Figure 8: Ablation study—an input image (left) accompanied by a set of images (right) generated using Stable Diffusion [Rombach et al. 2022] from a set of random input noises and conditioned by: (a) jointly optimized CLIP features c and initial noise ϵ, (b) CLIP features c optimized directly without the use of token regularization via CLIP model [Radford et al. 2021], (c) CLIP features c optimized using token regularization but without data augmentation (flipping, translation, and scaling), (d) our full approach. Note, how our full-fledged optimization better captures overall high level semantics. Source images: Adobe Stock.

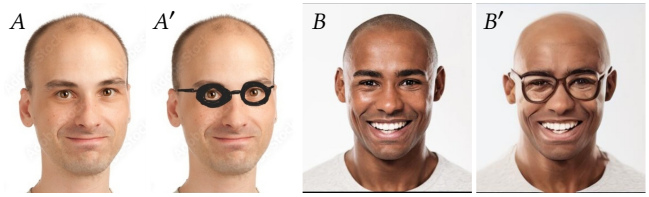

Figure 9: Extension—our approach can also be used in a specific scenario where the user draws a rough sketch A' over the input photo A to express the intended edit. In this case making the person to wear glasses. Then a different photo can be provided as a target image B to which this edit is applied B'. Source images: Adobe Stock.

scale σ, a built-in parameter of the diffusion process $\mathbf{U}^+$ (c.f. [Rombach et al. 2022]). In a typical workflow the user first manipulates σ to tune the extent to which the target image B is modified and then the analogy strength λ is set to express the influence of prescribed analogy $A : A'$ (see Fig. 11 and also our supplementary material for further examples of λ manipulation and video for live editing sessions).

In Figures 1, 3, and 4 we present various analogies performed by our approach. Note, how the input pair of images $A : A'$ can be highly diverse. There is no need to have objects aligned or be in a similar scale. The desired analogy is distilled automatically from the estimated CLIP features. Note also how our method preserves the structure of the target image and automatically respects higher level context (see, e.g., Fig. 3f where even though the image A' depicts plastic toy locomotives; in the output, we see a plastic toy car that does not resemble a locomotive).

4.1 Comparison

To our best knowledge our approach is the first technique that can perform high level image analogies on natural images. Although we can compare our method with the original image analogies [Hertzmann et al. 2001] it is obvious that the approach of Hertzmann et al. could not deliver comparable results to ours since they lack the ability to deduce high level semantics, and assume aligned image pair $A : A'$ as an input. Even if we provide an aligned pair it is still apparent (see Fig. 5) that the original image analogies only capture low level textural details and thus are unable to handle high level semantics without additional guidance.

Similar situation occurs when we compare our approach with the state-of-the-art in neural style transfer. Here we consider Deep Image Analogies [Liao et al. 2017] as a baseline solution, nevertheless, since the approach of Liao et al. does not support full image analogies (only images A and B can be specified) we can compare both techniques only in a specific scenario when $A = B$. Even in this limited setting it is visible (see Fig. 6) that high level semantics is difficult to capture using Deep Image Analogies.

Figure 10: Limitations—as our method was primarily designed to transfer high level features specified by the analogy $A : A'$ and preserve the overall structure of the target image B it may encounter difficulties when the user's intent is to perform pixel level stylization. As a future work we plan to combine our approach with previous image analogies approaches that can handle low level features better [Fišer et al. 2016; Texler et al. 2020a]. Source images: Adobe Stock.

To overcome the lack of comparable approaches to our method we developed a synthetic baseline solution that one can implement using off-the-shelf tools. We use BLIP [Li et al. 2022] to estimate text prompts for all input images A, A', and B. From those we then compute CLIP features, produce the desired analogy $c_{B'} = c_B + \lambda \cdot e_{A:A'}$, and finally we run standard Stable Diffusion in the `img2img` mode [Rombach et al. 2022]. Here the image B is first distorted by a small amount of Gaussian noise and then the diffusion process U^+ is executed while CLIP features $c_{B'}$ are used to guide the diffusion. Meng et al. [2022] use similar technique without CLIP guidance. The output of this approach is visible in Fig. 7 (`img2img`). Although the result follows the structure of the original image B that is already baked in this slightly noised input the analogy derived from the estimated text prompts does not capture semantics properly (see Fig. 12 and also our supplementary material for additional examples of this failure). When image variations mode of Stable Diffusion [LambdaLabsML 2022] is used, see Fig. 7 (`variation`) where the diffusion is guided directly using CLIP embedding of the input images (i.e., no optimization over text tokens is performed) the analogy is expressed better, however, the overall pose does not resemble the original B.

4.2 Ablation Study

To validate the necessity of all components in our proposed approach we performed a series of ablation studies. In Fig. 8 we demonstrate the importance of having CLIP features c and the initial noise ϵ optimized independently (8a). We also demonstrate the need for token-based regularization of CLIP features c (8b), the necessity of data augmentation (8c), and that to express the analogy the subtraction needs to be computed in the space of CLIP features c since the semantically meaningful ordering of tokens is unknown (see Fig. 13). In Fig. 14 we demonstrate how the proposed scheme that limits the number of tokens has positive impact on the ability to distill high level semantics.

5 LIMITATIONS AND FUTURE WORK

In our experiments, we demonstrate the benefits of enhancing image analogies framework with the capabilities of large language models and diffusion networks. These allow the transition of high-level features, which has been difficult to achieve using previous approaches. Nevertheless, there are still some limitations which we envision to be addressed in future work.

One of the limiting factors of our method is the fact that the user may find difficult to fine tune the analogy selectively in order to highlight specific details. Although it is possible to experiment with the analogy and guidance strength parameters the control over high level features being transferred to the target image is limited (see Fig. 15). As a future work we envision the users may have a possibility to specify their intent more closely, e.g., by sketching (see an example of initial prototype of this extension in Fig. 9), providing additional text prompts or by specifying multiple input exemplars that will better describe the intended analogy.

Since in our design we focused mainly on the overall structure and high level features, our approach may encounter difficulties when the analogy is focused solely on low level details (see Fig. 10). As a future work we envision to combine capabilities of our approach with traditional image analogies techniques [Fišer et al. 2016; Texler et al. 2020a] in order to properly capture high level context as well as being able to faithfully reproduce pixel level details.

6 CONCLUSION

We presented an approach to image analogies that in contrast to previous techniques focused on low level textual details respects semantics of the specified analogy and is able transfer high level features while preserving the structure of the target image. A key component that enables us to achieve such an upgrade of the original image analogies framework [Hertzmann et al. 2001] is the connection of large language model [Radford et al. 2021] with the power of diffusion networks [Rombach et al. 2022]. Although a similar combination was already used in previous works, in our approach we demonstrate that one can leverage the power of large language model without the need to work with text prompts explicitly. We believe that our work inspires future exploration of possible extensions to a set of controls available to the users who work with generative models.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their valuable feedback. This research was supported by Adobe, the Research Center for Informatics, grant No. CZ.02.1.01/0.0/0.0/16 019/0000765, and by the Grant Agency of the Czech Technical University in Prague, grants No. SGS22/173/OHK3/3T/13 and No. SGS23/173/OHK3/3T/13.

REFERENCES

Michael Ashikhmin. 2001. Synthesizing Natural Textures. In *Proceedings of Symposium on Interactive 3D graphics*. 217–226.

Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended Diffusion for Text-driven Editing of Natural Images. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 18208–18209.

Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei A Efros. 2022. Visual Prompting via Image Inpainting. In *Advances in Neural Information Processing Systems*.

Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. 2009. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics* 28, 3 (2009), 24.

Pierre Bénard, Forrester Cole, Michael Kass, Igor Mordatch, James Hegarty, Martin Sebastian Senn, Kurt Fleischer, Davide Pesare, and Katherine Breeden. 2013. Stylizing Animation By Example. *ACM Transactions on Graphics* 32, 4 (2013), 119.

Albert S Berahas, Jorge Nocedal, and Martin Takac. 2016. A Multi-Batch L-BFGS Method for Machine Learning. In *Advances in Neural Information Processing Systems*.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023. InstructPix2Pix: Learning to Follow Image Editing Instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*. 1877–1901.

Mathilde Caron, Hugo Touvron, Ishan Misra, Herve Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of IEEE International Conference on Computer Vision*. 9650–9660.

Olga Diamanti, Connelly Barnes, Sylvain Paris, Eli Shechtman, and Olga Sorkine-Hornung. 2015. Synthesis of Complex Image Appearance from Limited Exemplars. *ACM Transactions on Graphics* 34, 2 (2015), 22.

Jakub Fišer, Ondřej Jamriška, Michal Lukáč, Eli Shechtman, Paul Asente, Jingwan Lu, and Daniel Sýkora. 2016. StyLit: Illumination-Guided Example-Based Stylization of 3D Renderings. *ACM Transactions on Graphics* 35, 4 (2016), 92.

Jakub Fišer, Ondřej Jamriška, David Simons, Eli Shechtman, Jingwan Lu, Paul Asente, Michal Lukáč, and Daniel Sýkora. 2017. Example-Based Synthesis of Stylized Facial Animations. *ACM Transactions on Graphics* 36, 4 (2017), 155.

William T. Freeman, Thouis R. Jones, and Egon C. Pasztor. 2002. Example-Based Super-Resolution. *IEEE Computer Graphics and Applications* 22, 2 (2002), 56–65.

David Futschik, Menglei Chai, Chen Cao, Chongyang Ma, Aleksei Stoliar, Sergey Korolev, Sergey Tulyakov, Michal Kučera, and Daniel Sýkora. 2019. Real-Time Patch-Based Stylization of Portraits Using Generative Adversarial Network. In *Proceedings of the ACM/EG Expressive Symposium*. 33–42.

David Futschik, Michal Kučera, Michal Lukáč, Zhaowen Wang, Eli Shechtman, and Daniel Sýkora. 2021. STALP: Style Transfer with Auxiliary Limited Pairing. *Computer Graphics Forum* 40, 2 (2021), 563–573.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. 2022a. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *arXiv:2208.01618*.

Rinon Gal, Or Patashnik, Haggai Maron, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. 2022b. StyleGAN-NADA: CLIP-Guided Domain Adaptation of Image Generators. *ACM Transactions on Graphics* 41, 4 (2022), 141.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2016. Image Style Transfer Using Convolutional Neural Networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 2414–2423.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. In *arXiv:2208.01626*.

Aaron Hertzmann, Charles E. Jacobs, Nuria Oliver, Brian Curless, and David H. Salesin. 2001. Image Analogies. In *SIGGRAPH Conference Proceedings*. 327–340.

Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. 2022. Cascaded Diffusion Models for High Fidelity Image Generation. *Journal of Machine Learning Research* 23, 47 (2022), 1–33.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 5967–5976.

Ondřej Jamriška, Šárka Sochorová, Ondřej Texler, Michal Lukáč, Jakub Fišer, Jingwan Lu, Eli Shechtman, and Daniel Sýkora. 2019. Stylizing Video by Example. *ACM Transactions on Graphics* 38, 4 (2019), 107.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-Based Real Image Editing with Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Nicholas I. Kolkin, Jason Salavon, and Gregory Shakhnarovich. 2019. Style Transfer by Relaxed Optimal Transport and Self-Similarity. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 10051–10060.

Gihyun Kwon and Jong Chul Ye. 2023. Diffusion-based Image Translation using disentangled style and content representation. In *Proceedings of International Conference on Learning Representations*.

LambdaLabsML. 2022. Lambda Diffusers - Stable Diffusion Image Variations. https://github.com/LambdaLabsML/lambda-diffusers

Chuan Li and Michael Wand. 2016. Combining Markov Random Fields and Convolutional Neural Networks for Image Synthesis. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 2479–2486.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *International Conference on Machine Learning*. 12888–12900.

Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. 2017. Universal Style Transfer via Feature Transforms. In *Advances in Neural Information Processing Systems*. 385–395.

Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. 2017. Visual Attribute Transfer Through Deep Image Analogy. *ACM Transactions on Graphics* 36, 4 (2017), 120.

Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. 2019. Few-Shot Unsupervised Image-to-Image Translation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 10551–10560.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2022. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *Proceedings of International Conference on Learning Representations*.

Tomáš Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of International Conference on Learning Representations*.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Null-text Inversion for Editing Real Images using Guided Diffusion Models. In *arXiv:2211.09794*.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *Proceedings of International Conference on Machine Learning*. 16784–16804.

Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. 2020. Contrastive Learning for Unpaired Image-to-Image Translation. In *Proceedings of European Conference on Computer Vision*.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. In *Proceedings of IEEE International Conference on Computer Vision*. 2085–2094.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*. 8748–8763.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. (2019).

Aditya Ramesh. 2022. How DALL·E 2 Works. http://adityaramesh.com/posts/dalle2/dalle2.html

Lincoln Ritter, Wilmot Li, Brian Curless, Maneesh Agrawala, and David Salesin. 2006. Painting With Texture. In *Proceedings of Eurographics Symposium on Rendering*. 371–376.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. 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*.

Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014).

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *Proceedings of International Conference on Learning Representations*.

Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. 2022. ZeroCap: Zero-Shot Image-to-Text Generation for Visual-Semantic Arithmetic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17897–17907.

Ondřej Texler, David Futschik, Jakub Fišer, Michal Lukáč, Jingwan Lu, Eli Shechtman, and Daniel Sýkora. 2020a. Arbitrary Style Transfer Using Neurally-Guided Patch-Based Synthesis. *Computers & Graphics* 87 (2020), 62–71.

Ondřej Texler, David Futschik, Michal Kučera, Ondřej Jamriška, Šárka Sochorová, Menglei Chai, Sergey Tulyakov, and Daniel Sýkora. 2020b. Interactive Video Stylization Using Few-Shot Patch-Based Training. *ACM Transactions on Graphics* 39, 4 (2020), 73.

Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. 2022. Splicing ViT Features for Semantic Appearance Transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10748–10757.

Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. 2020. Cross-domain Correspondence Learning for Exemplar-based Image Translation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 5143–5153.

Yang Zhou, Huajie Shi, Dani Lischinski, Minglun Gong, Johannes Kopf, and Hui Huang. 2017. Analysis and Controlled Synthesis of Inhomogeneous Textures. *Computer Graphics Forum* 36, 2 (2017), 199–212.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *Proceedings of IEEE International Conference on Computer Vision*. 2242–2251.

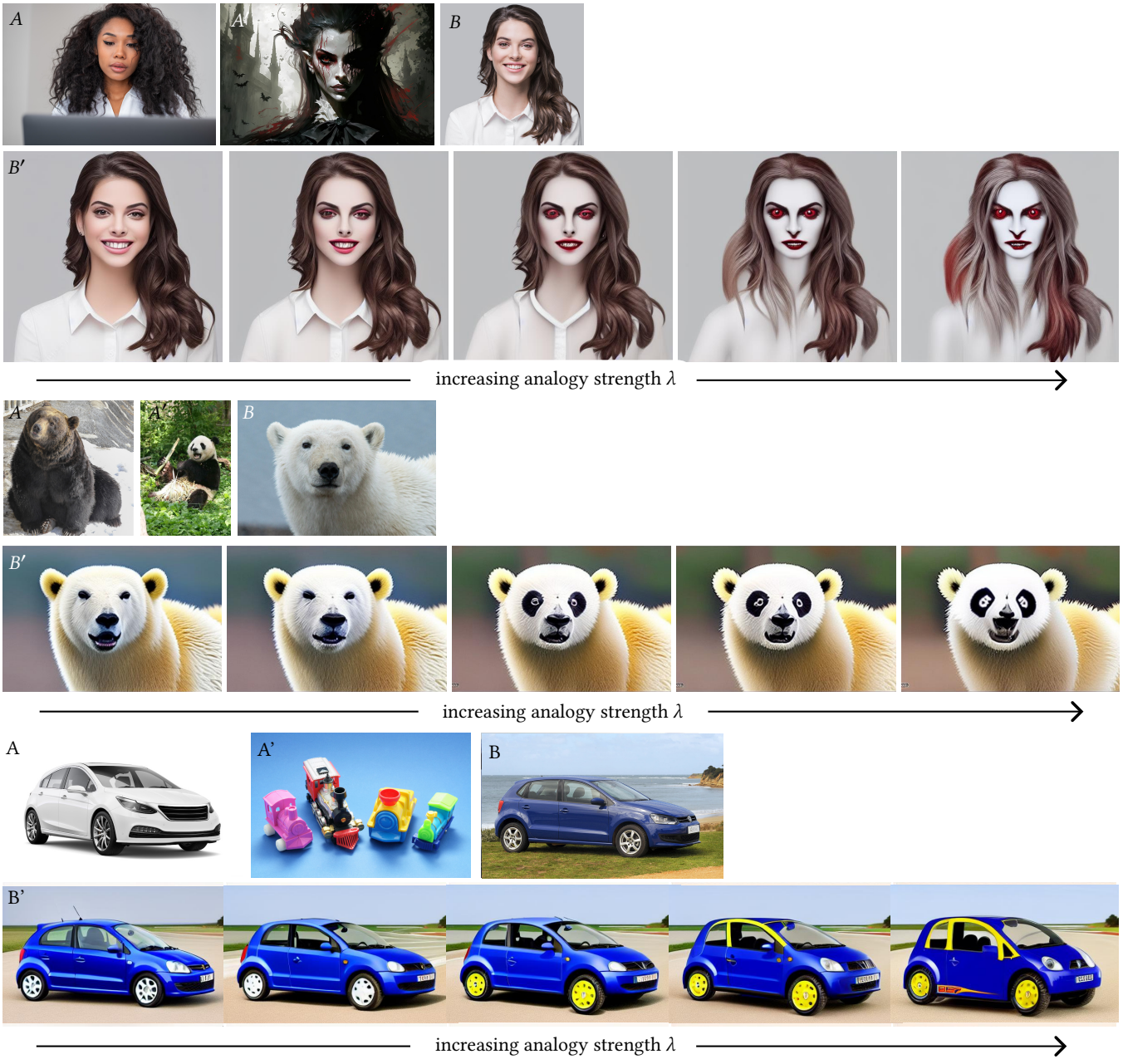

Figure 11: Examples of diffusion image analogies $A : A' :: B : B'$ produced using our approach with gradually increasing analogy strength λ. Note, how increasing λ makes the prescribed analogy more apparent. Source images: © Artanisen (Brown Bear A), © Cliff (Panda A'), © The NRMA (Volkswagen Golf B), Adobe Stock the rest.

Figure 12: Comparison with Stable Diffusion [Rombach et al. 2022] (cont.): A ="*a black and white photo of a tree in a field*", A' ="*a lone tree in a field of green grass*", and B ="*a black and white photo of a man with a long beard*". **Note, how in the estimated description of A' color is not explicit mentioned and thus also not pronounced well in the resulting analogy B'. Source images: © Subhamshome28 B, Adobe Stock the rest.**

Figure 13: Ablation study—a comparison of two analogies performed using our approach vs. a scenario where instead of computing the difference of estimated CLIP features c, token vectors $\mathcal{K}$ are subtracted directly. Since the token order is crucial for this operation it is unclear which tokens need to be subtracted. In the space of CLIP features c such an ordering issue is not present and thus the analogy is accurate. Source images: Adobe Stock.

Figure 14: Ablation study—CLIP features c for an input image (left) are estimated without the regularization on the number of tokens (a) and with the regularization (b). Note, how the regularization distills semantics so that when the diffusion process U^+ is executed with different random noises the output better reassembles the original setting (apples stacked on top of each other) whereas without regularization we can see other objects in the generated images. Source image: Adobe Stock.

Figure 15: Limitations—when specifying the analogy $A : A'$ the original user's intent was to add a mountain to the horizon in the image B. However, since in the image A' besides the mountain contains also a lake its transfer to the image B is performed as well although it was not originally intended. As a future work we envision to employ additional more specific control over the analogy. Source images: © Martin Falbisoner A, Adobe Stock the rest.