

Cross-Image Attention for Zero-Shot Appearance Transfer

Yuval Alaluf*

Daniel Garibi*

Or Patashnik

Hadar Averbuch-Elor

Daniel Cohen-Or

Tel Aviv University

<https://garibida.github.io/cross-image-attention/>

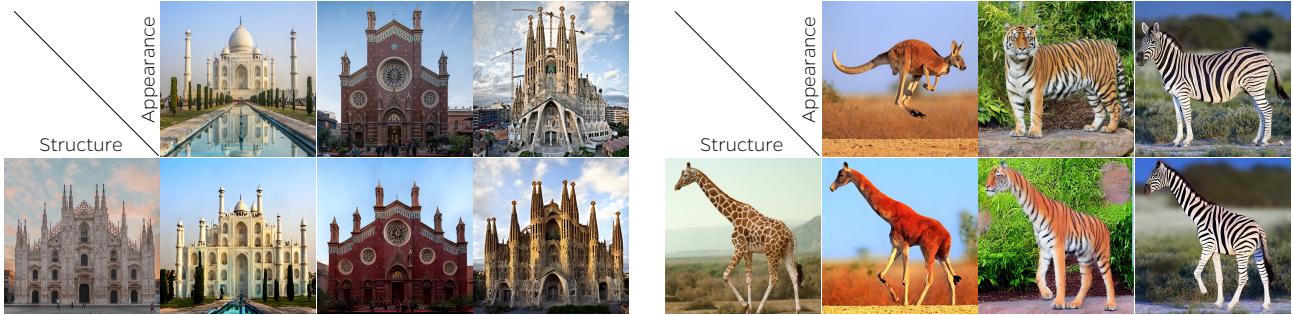


Figure 1. Given two images depicting a source structure and a target appearance, our method generates an image merging the structure of one image with the appearance of the other. We do so in a *zero-shot* manner, with no optimization or model training required while supporting appearance transfer across images that may differ in size and shape.

Abstract

Recent advancements in text-to-image generative models have demonstrated a remarkable ability to capture a deep semantic understanding of images. In this work, we leverage this semantic knowledge to transfer the visual appearance between objects that share similar semantics but may differ significantly in shape. To achieve this, we build upon the self-attention layers of these generative models and introduce a cross-image attention mechanism that implicitly establishes semantic correspondences across images. Specifically, given a pair of images — one depicting the target structure and the other specifying the desired appearance — our cross-image attention combines the queries corresponding to the structure image with the keys and values of the appearance image. This operation, when applied during the denoising process, leverages the established semantic correspondences to generate an image combining the desired structure and appearance. In addition, to improve the output image quality, we harness three mechanisms that either manipulate the noisy latent codes or the model’s internal representations throughout the denoising process. Importantly, our approach is zero-shot, requiring no optimization or training. Experiments show that our method is effective across a wide range of object categories and is robust to variations in shape, size, and viewpoint between the two input images.

1. Introduction

The rapid growth and adoption of powerful generative models have granted users an unprecedented level of freedom to create stunning, diverse visual content with relative ease [4, 17, 52, 60, 63, 65, 67]. In parallel with these advancements in generative capabilities, many have sought new avenues to gain greater control over the *manipulation* of visual content using these generative models.

In this work, we explore image manipulation within the context of appearance transfer, where we aim to transfer the visual appearance of a concept from one image to a concept present in another image. Consider, for example, transferring the appearance of a zebra to a giraffe (see Figure 1). Successfully accomplishing this task requires first associating semantically similar regions between the giraffe and zebra (e.g., their legs, head, and neck) and then transferring the zebra’s appearance in a realistic manner through these associations without altering the structure of the giraffe. Furthermore, a particular challenge in this task is establishing these associations across images containing objects from different categories that vary in shape, as well as images with differing viewpoints and illuminations. Previous attempts assume that appearance transfer is performed between objects of similar shape [19, 51, 74], or require training a model for a specific class of objects [56, 81].

* Denotes equal contribution

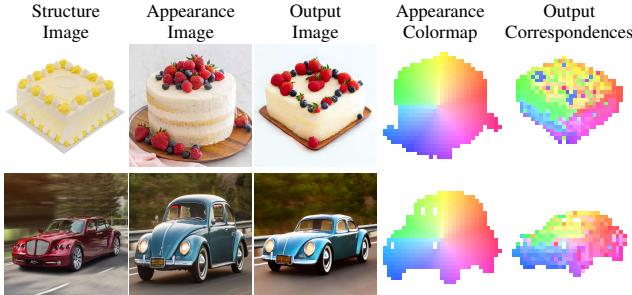


Figure 2. Implicitly finding correspondences via our cross-image attention applied between two images. For each pixel in the structure image, we identify the pixel in the appearance image that attains the highest activation in its cross-image attention map. The output correspondences represent the pixel mapping from the structure to the appearance image using the maximum activations. As shown, these correspondences are semantically aligned (e.g., matching the fruits on the cake and the bodies of the cars).

Analyzing the inner workings of recent large-scale diffusion models, many works have demonstrated that the cross- and self-attention mechanisms of the denoising network implicitly encode strong semantic information from the generated image [9, 23, 26, 58, 73, 75]. Building on the functions of the queries, keys, and values within these self-attention layers, our key insight is to employ the self-attention mechanism across *different* images, which we term *Cross-Image Attention*. As illustrated in Figure 2, when applied to images featuring distinct subjects with varying shapes and structures, this cross-image attention forms strong associations between similar semantic regions in the two images.

More specifically, given an appearance image and a structure image, we begin by inverting the two images into the latent space of a pretrained text-to-image diffusion model [63]. Then, at each timestep of the denoising process, we compute a modified self-attention map by multiplying the *queries* corresponding to the structure image with the *keys* of the appearance image. This cross-image operation establishes implicit semantic correspondences between the two images, without requiring additional supervision, as illustrated in Figure 2. Then, by multiplying the resulting cross-image attention map by the *values* of the appearance image, we can accurately transfer each pixel from the appearance image to the corresponding, semantically similar pixel(s) in the structure image.

While the cross-image mechanism is conceptually simple, we observe that it alone is not sufficient for attaining an accurate semantic transfer between the two images, often leading to noticeable artifacts in the resulting image. We attribute these artifacts to the domain gap between the queries of the structure image and the keys and values of the appearance image. To address this challenge, we employ three mechanisms aimed at enhancing transfer quality. First, we amplify the variance of the cross-image at-

tention maps, making them more focused on capturing only the most semantically similar image regions. Second, we adapt the classifier-free guidance technique [27, 52] to the task of appearance transfer and strengthen the influence of our cross-image attention operation on the generated image during the denoising process. Finally, we leverage the AdaIN [30] mechanism to align the image statistics of the appearance and output images, better preserving the color of the appearance image.

We illustrate the versatility of our cross-image attention and show its effectiveness for zero-shot appearance transfer across a wide range of object domains. This includes challenging image pairs containing objects with substantial variations in shape, viewpoint, and number of instances. We also perform quantitative comparisons to existing methods, further demonstrating that our results better capture the target appearance while preserving the source structure.

2. Related Works

Appearance Transfer The task of appearance transfer can be seen as a specialized form of image-to-image translation. However, unlike Neural Style Transfer [21, 34, 39, 54, 78], which focuses on transferring a *global* artistic style across images, our focus is on *semantic style transfer*, where we aim to transfer the appearance between semantically related regions in two images.

Early generative-based approaches trained a Generative Adversarial Network (GAN)[24] on a large collection of either paired[32] or unpaired images [35, 46, 55, 77, 81]. Notably, Park *et al.* [56] introduced Swapping Autoencoders, where they train an autoencoder, separately encoding the structure and the appearance of an image. Then, to transfer the appearance from one image to another, they take the structure representation from one image and the appearance code from the other and pass them together to the decoder. However, this approach necessitates training a dedicated generator for each target domain (e.g., churches or cats) and requires collecting a large domain-specific dataset.

To reduce the level of supervision required, several methods learn a mapping using a single exemplar [7, 15, 43, 45, 71]. Tumanyan *et al.* [74] train a dedicated generator for a single image pair and utilize a pretrained DINOViT [10, 18] to extract structure and style information from input images, injecting them into the training process to guide the transformation. This approach, however, requires training a dedicated generator for each pair of images, which takes dozens of minutes per input. Moreover, the technique mainly works well between images with relatively similar shapes.

Most similar to our approach, recent works have sought to leverage powerful large-scale diffusion models for appearance transfer without additional inputs or model training [19, 40, 51]. These methods typically incorporate losses

applied over the noisy latent codes to guide the denoising process toward generating images depicting the structure of a given image while adapting its appearance. However, unlike our method, the appearance losses in these works rely on a global appearance descriptor and do not consider the semantic correspondences between the images. As a result, these methods are often limited to a coarse appearance transfer or constrained to transferring appearance between objects of the same category, size, and shape.

In contrast to the above approaches, our method operates with no training or per-image optimization and respects the semantic correspondences between the two images when transferring appearance. Furthermore, our method requires only a single forward pass through a pretrained diffusion model and is applicable to diverse image pairs that may contain cross-domain subjects.

Semantic Correspondences with Generative Models

The task of finding correspondences between two images has been a longstanding challenge in computer vision, ranging from classical approaches [6, 47, 64] to learning-based techniques [1, 12, 28, 50, 53, 62, 66, 70]. With the rapid improvements in generative models, many have explored adapting these models for the task of semantic correspondence. Peebles *et al.* [59] leverage the latent space of a pretrained GAN to train a Spatial Transformer [33] tasked with densely aligning a set of images from a specific domain. In the context of diffusion models, numerous works have demonstrated that the intermediate features of the denoising network of a pretrained diffusion model [63] can effectively establish semantic correspondences across different object categories [25, 48, 73, 79].

Image Editing with Diffusion Models Building on the recent advancements in large-scale diffusion models [52, 60, 63, 65, 67], many works have explored new avenues to gain more precise control over the generative process [3, 5, 29, 42, 52, 76, 80], further utilizing these models for various downstream editing tasks [2, 8, 13, 16, 36, 49]. To provide users with additional control over the generation and editing process, recent works have also utilized user-provided spatial conditions to specify the region that should be altered [2, 3, 5, 16, 42, 76, 80]. Notably, numerous works have shown that manipulating the internal representations of the denoising network, particularly its attention layers, is effective for image editing [9, 19, 22, 23, 26, 44, 57, 58, 75], as well as for finer control over the image generation process [11, 41, 61].

Recently, MasaCtrl [9] demonstrated that freezing the keys and values of the self-attention layers when performing non-rigid edits of an image more faithfully preserves the image’s original appearance. TokenFlow [23] and Infusion [37] extend this technique to preserve the appearance of different frames when editing a video. In our method,

we also inject keys and values into the self-attention layers. However, unlike the methods mentioned above, this injection is performed between *different* images rather than between an image and its output edit.

3. Method

Given a pair of images (I^{struct}, I^{app}), we wish to generate an output image I^{out} depicting the structure of the subject present in I^{struct} with the appearance of the subject in I^{app} . To perform the transfer, we utilize a pretrained text-to-image diffusion model, namely Stable Diffusion [63]. We first briefly review concepts related to the self-attention layers within image diffusion models. We then introduce our Cross-Image Attention mechanism, which is the core of our proposed method, and then describe how it can be utilized for zero-shot appearance transfer.

3.1. Preliminaries

We begin by describing the self-attention layers that compose the denoising U-Net within the image diffusion model. At each timestep t of the denoising process, the noised latent code z_t is fed as input to the denoising network. Consider a specific self-attention layer ℓ . The intermediate features of the network at ℓ , denoted $\phi_\ell(z_t)$, are first projected into queries $Q = f_Q(\phi(z_t))$, keys $K = f_K(\phi(z_t))$, and values $V = f_V(\phi(z_t))$ through learned linear projections f_Q, f_K, f_V .

For each query vector $q_{i,j}$ located at spatial location (i,j) of Q , we calculate a similarity score, or attention score, with respect to all keys in K , reflecting how relevant each key is to the corresponding query. These attention scores are then normalized using a softmax operation, defining the weight each value will have when updating the features at position (i,j) . Finally, the weighted values are aggregated to produce the output for each query position. Formally, we compute:

$$A_{(i,j)} = \text{softmax} \left(\frac{q_{i,j} \cdot K^T}{\sqrt{d}} \right) \quad (1)$$

$$\Delta\phi_{(i,j)} = A_{(i,j)} \cdot V,$$

where $A_{(i,j)}$ represents the attention map at (i,j) and $\Delta\phi_{(i,j)}$ denotes the aggregated output feature at (i,j) used to update the spatial features $\phi(z_t)$. This process is applied independently for all queries, enabling the model to capture correspondences across the entire image.

3.2. Cross-Image Attention

In a recent work, Cao *et al.* [9] explored the self-attention layers within the denoising network of a text-to-image diffusion model. They show that keeping the keys and values of these self-attention layers fixed aids in preserving the visual characteristics of objects when applying non-rigid manipulations over a given image.

Building on this insight, in this work we demonstrate the significant roles played by the queries, keys, and values in encoding semantic information of the generated image. More specifically, we observe that one can utilize the queries, keys, and values from these self-attention layers to transfer semantic information between *different* images. As shall be shown, the *queries* determine the semantic meaning of each spatial location. Next, the *keys* offer context for each query, allowing the model to weigh the importance of different parts of the image for a specific query position. Lastly, the values represent the content we aim to generate and define the information that will be used for determining the features of each query position.

To define our cross-image attention layer, we replace the keys and values corresponding to one image with the keys and values of another image. We demonstrate that by doing so, it is possible to implicitly transfer the visual appearance between semantically similar objects in the images. More precisely, we replace the keys and values corresponding to the output image I^{out} with the keys and values corresponding to the appearance image I^{app} . Formally, the output of our cross-image attention layer is given by:

$$\Delta\phi^{\text{cross}} = \text{softmax} \left(\frac{Q_{out} \cdot K_{app}^T}{\sqrt{d}} \right) V_{app}. \quad (2)$$

We illustrate the roles of the keys and queries in Figure 3. In each column, we highlight one of three query locations marked by red, yellow, and green circles. For each row, we then display the attention maps obtained for each query location using different combinations of queries and keys corresponding to I^{struct} and I^{app} . As shown, when multiplying the keys and queries corresponding to the same image (i.e., computing $Q_{struct} \cdot K_{struct}^T$ or $Q_{app} \cdot K_{app}^T$), each query attends to semantically similar regions within the image. For example, consider the yellow query. In the attention of the structure image (row one), the query attends to the legs of the giraffe as it is located on the giraffe's leg. Conversely, in the appearance image (row two), the yellow query lies on the grass of the image background, and hence, the query attends to nearby grass-like pixels in the image.

In the bottom row, we apply our cross-image attention mechanism and compute $Q_{struct} \cdot K_{app}^T$. As shown, the queries now attend to a semantically corresponding region in the zebra image. For example, the red query now attends to the head of the zebra while the yellow query attends to the leg of the zebra. Interestingly, these associations are established even though the two animals differ significantly in shape. Consequently, by multiplying the resulting attention maps with the values V_{app} from the appearance image, we can accurately project semantically similar regions from the zebra image onto the giraffe image. This enables one to transfer the zebra's appearance onto the giraffe's structure.

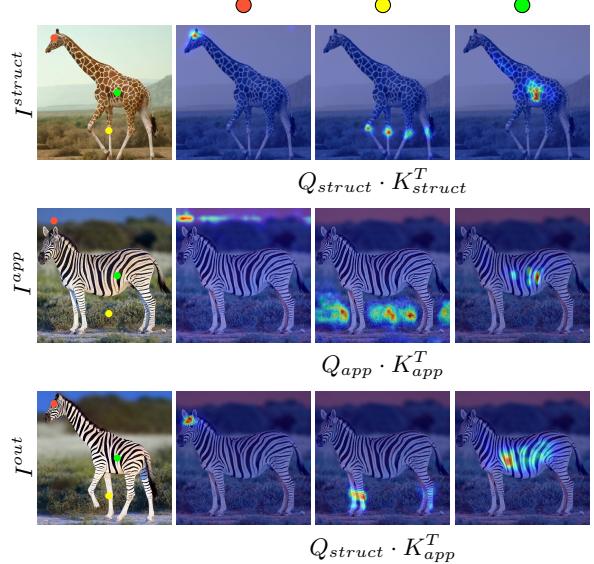


Figure 3. Establishing correspondences through the keys and queries of self-attention and cross-image attention. Using color-coded markers, we denote three queries corresponding to different semantic regions of the structure image (the giraffe's head, leg, and body). These markers are placed in the same pixel locations in the three images. In each row, we illustrate the attention maps obtained by computing various dot products of the queries and keys of the two images computed at a single layer. In the first two rows, we show the self-attention maps obtained using queries and keys originating from the same image, resulting in each query focusing on semantically similar regions in the image. For instance, the yellow query attends to the legs of the giraffe in the structure image and to nearby grass pixels in the background of the appearance image. In the bottom row, we use our cross-image attention, aligning the queries Q_{struct} with the keys K_{app} . In doing so, each query on the giraffe corresponds to semantically similar regions of the zebra. For example, the red query attends to the head of the zebra while the yellow query points to its legs.

3.3. Appearance Transfer

We now turn to describe how our cross-image attention mechanism can be utilized for semantic-based appearance transfer, as depicted in Figure 4. Given input images I^{struct} and I^{app} , we begin by inverting them using the edit-friendly DDPM inversion introduced in Huberman *et al.* [31]. After obtaining the inverted latents, denoted by z_T^{struct} and z_T^{app} , we perform a denoising process along two parallel paths, resulting in the reconstruction of I^{app} and the generation of I^{out} . To initialize this denoising process, we set the latent z_T^{out} representing our output image to be equal to z_T^{struct} .

At each timestep t , we pass the two latent codes z_t^{out} and z_t^{app} to the denoising U-Net model. Within the decoder of the U-Net, we replace the standard self-attention with our cross-image attention and compute the modified output using Equation (2). In practice, our cross-image attention replaces the standard self-attention layers in the U-Net decoder layers with output resolutions of 32×32 and 64×64 .

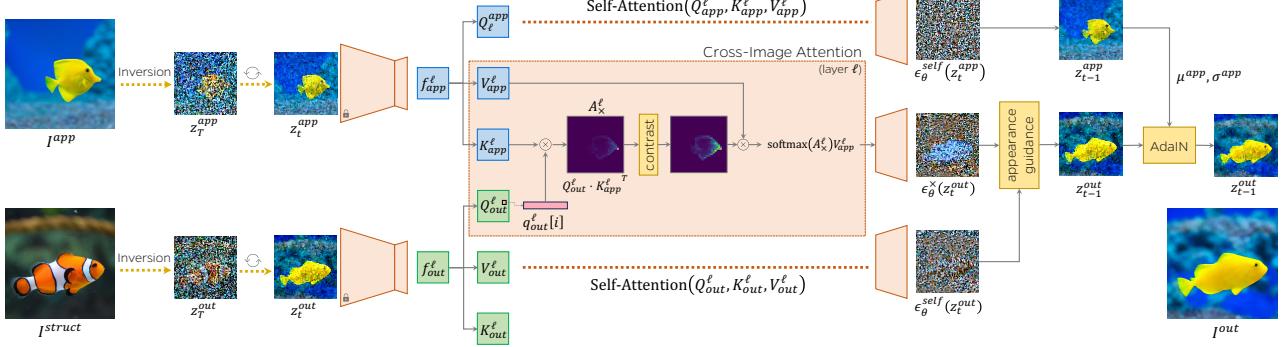


Figure 4. Method overview. Given I^{struct} and I^{app} , we begin by inverting the two images into the latent space of a pretrained image diffusion model, resulting in latents z_T^{struct} and z_T^{app} . To initialize our output latent, we set $z_T^{out} = z_T^{struct}$. Consider some timestep t and self-attention layer ℓ . To compute the next latent z_{t-1}^{out} , we compute our cross-image attention map defined in Equation (2) by mixing the keys and values from z_t^{app} with the query of z_t^{out} . To improve the output image quality, we introduce three extensions. First, we apply a contrast operation over the cross-image attention map, encouraging the Q_{out} to attend to a smaller set of keys in K_{app} . Next, we introduce an appearance guidance mechanism akin to classifier-free guidance used for text-guided image synthesis. Finally, we apply an AdaIN operation over z_{t-1}^{out} to better align with the feature statistics of z_{t-1}^{app} . This process is repeated across multiple timesteps of the denoising process and across multiple layers of the network decoder, resulting in the gradual appearance transfer from I^{app} to I^{struct} .

While this simple injection mechanism allows for the transfer of pixels from the appearance image to semantically similar regions in the structure image, it may still result in noticeable artifacts in the generated output image. We attribute this issue to the presence of a domain gap between the queries, keys, and values that are computed from latent codes of two distinct images, resulting in a lower-quality output image. To improve the output image quality, we introduce several additional mechanisms to guide the appearance transfer, detailed below.

Attention Map Contrasting First, we observe that some queries of z_t^{out} attain a high similarity to many keys of z_t^{app} . This can be observed in Figure 5 where the cross-image attention map obtained by our method returns a sparse and unfocused attention map (fourth column). This is in contrast to the attention map obtained by the standard self-attention mechanism, which attains high similarities in a concentrated region of the image (third column). These sparse attention maps may lead to inaccurate transfers because the output value of each query is computed using an aggregation of pixels spanning many different image regions. This, in turn, can result in unwanted artifacts in the final image.

To encourage the attention maps to focus on more concentrated regions in the image, we apply a contrast operation to increase the variance of the attention maps. Given A_x^ℓ obtained from our cross-image operation, we update

$$A_x^\ell \leftarrow (A_x^\ell - \mu(A_x^\ell))\beta + \mu(A_x^\ell), \quad (3)$$

where μ is the mean operation and β is the contrast factor, empirically set to $\beta = 1.67$. Note that this operation is applied before the attention map is multiplied with V_{app} .

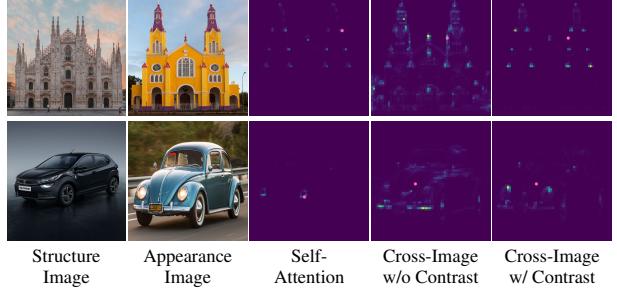


Figure 5. Attention map contrasting. When applying the standard self-attention, queries often attend to a small set of semantically similar pixels (third column). In contrast, our cross-image attention map may cause a specific query (highlighted in pink) to attain high activations across many pixels across the entire image (fourth column). By applying our contrast operation over the cross-image attention maps, the maps behave more similarly to the standard self-attention, focusing on the more semantically similar image regions (rightmost column).

Appearance Guidance Next, we adapt the concept of classifier-free guidance [27], which has been shown to improve the overall quality of generated images, to the realm of appearance transfer. At each denoising step t we perform two forward passes through the denoising network: (i) $\epsilon^\times = \epsilon_\theta^\times(z_t^{out})$ using our cross-image attention layer, and (ii) $\epsilon^{\text{self}} = \epsilon_\theta^{\text{self}}(z_t^{out})$ using the original self-attention layer of the network. Given the two noise predictions, we then define the final predicted noise ϵ^t as:

$$\epsilon^t = \epsilon^{\text{self}} + \alpha (\epsilon^\times - \epsilon^{\text{self}}), \quad (4)$$

where α is the guidance scale. The next latent code z_{t-1}^{out} is then sampled using the modified noise ϵ^t .



Figure 6. Semantic-based appearance transfer results obtained by our method. In each grid, the leftmost column displays the input structure images, while the topmost row presents the input appearance images. The remaining 3×3 grid showcases the results of the appearance transfer between each corresponding structure and appearance image.

Intuitively, this guidance mechanism shifts the noisy latent code towards denser regions of the distribution associated with the target appearance while moving away from the original appearance. By reaching denser regions of the distribution, we obtain more plausible images, resulting in fewer artifacts.

AdaIN In addition to the artifacts handled by the previously described mechanisms, we observe a shift in the color distribution between the output image and the input appearance image. To address this, we utilize the AdaIN operation [30], originally introduced for style transfer and known for effectively matching feature statistics between latent representations. We find that applying AdaIN on z_t^{out} with respect to z_t^{app} assists in gradually aligning the color distribution of the output and appearance images. Specifically, we update

$$z_t^{out} \leftarrow \text{AdaIN}(z_t^{out}, z_t^{app}), \quad (5)$$

where the statistics of z_t^{out} are adjusted to match those of z_t^{app} , assisting in aligning their color distributions.

However, we notice that the statistics computed by the AdaIN operation are sensitive to the size of the objects. As a result, AdaIN may not be effective when the objects depicted in the images significantly vary in size. To address this, we apply a mask over the latents z_t^{out} and z_t^{app} and restrict the AdaIN operation to compute the feature statistics only on a foreground mask containing the object. To create the object masks, we employ the unsupervised self-segmentation technique introduced in Patashnik *et al.* [58].

4. Experiments

In the following section, we demonstrate the effectiveness of our cross-image attention technique for the task of appearance transfer.



Figure 7. Cross-domain appearance transfer. Our approach can transfer appearance between cross-domain objects. This transfer is possible even in a zero-shot setting thanks to the strong correspondences already captured by the diffusion model itself.

4.1. Evaluations and Comparisons

Evaluation Setup We evaluate the performance of our cross-image attention mechanism in comparison to state-of-the-art semantic-based appearance transfer methods. These works range from methods that require training a generator for each target domain (Swapping Autoencoder [56]) or each input image pair (SpliceVIT [74]) to those relying on external models to guide an inference-time optimization process (e.g., DiffuseIT [40]). Results for all methods are produced using their official implementations and default parameters. Additional details can be found in Appendix A.

Qualitative Evaluation In Figure 6 we illustrate appearance transfer results obtained by our method across three object domains. As can be seen, our method is effective in transferring the visual appearance in a semantically faithful manner. This holds even for challenging image pairs that may contain variations in object shape. For example, in

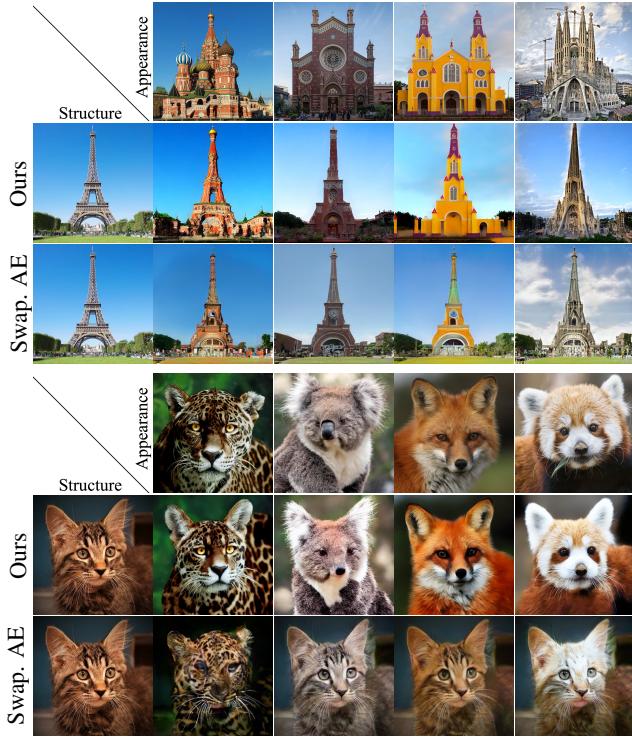


Figure 8. Comparison to Swapping Autoencoder (SA) [56]. We provide a comparison to SA using their pretrained churches and animal faces models. For each input structure image (shown to the left), we show transfer results obtained through four different appearance images (shown in the top row).

the leftmost grid, our method successfully transfers prominent features between the buildings such as the dome tops in the first building or the columns of the Taj Mahal in the second row. Moreover, in the middle example, we successfully transfer key visual features between the cars such as the headlights of the blue beetle (leftmost column) or the front grill of the red car (middle column).

Next, in Figure 7 we present more challenging cross-domain results where the structure and appearance images come from different object categories. Our method can still generate semantically plausible images, such as between the airplane and the hummingbird in the leftmost column. This transfer also works surprisingly well between objects with less shared semantics such as a watch and a phone or a shirt and a coffee mug. We do observe, however, that transfer between cross-domain images is generally more challenging due to the less accurate correspondences typically established by the model. For instance, in the fourth column, the tie of the tuxedo is not transferred to the output image.

Qualitative Comparison We now turn to qualitatively compare our cross-image attention mechanism to existing appearance transfer techniques. Since Swapping Autoencoder (SA) requires a dedicated generator for each domain, we begin with a comparison to SA using their pretrained

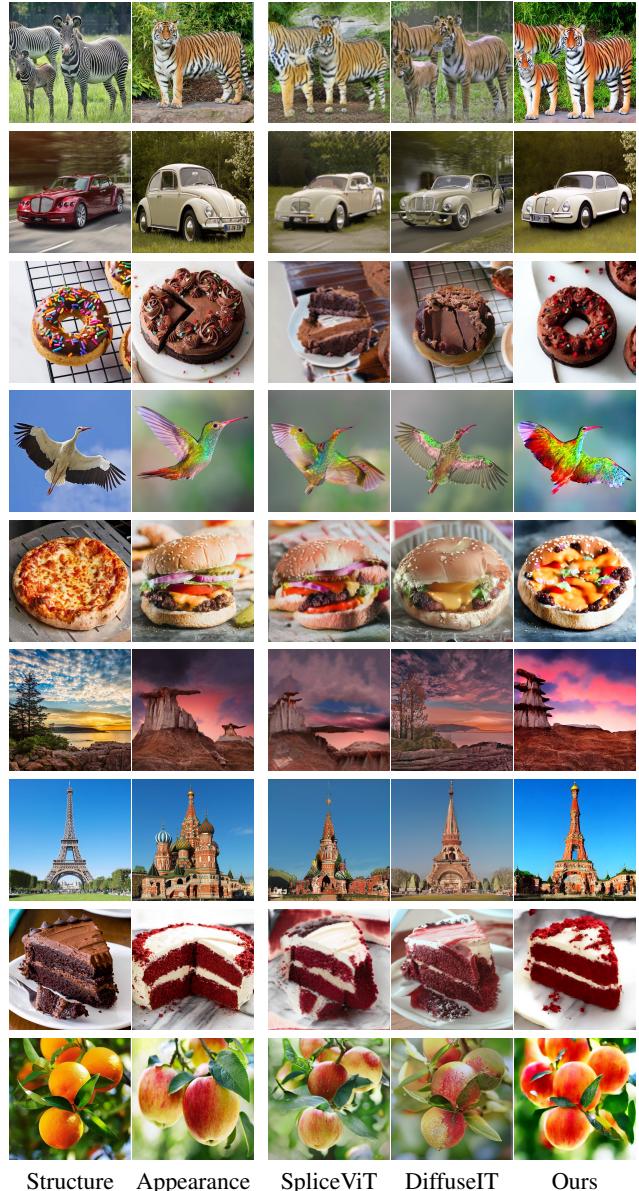


Figure 9. Qualitative comparison to additional appearance transfer techniques. In each row, we provide the input structure and appearance images, followed by the results obtained by each method.

church and AFHQ [14] models. Results are presented in Figure 8. SA effectively maintains the source structure while transferring the general color scheme of the target appearance. However, it often falls short of capturing the semantic details. For instance, it struggles to transfer the gold dome and colorful patterns in the leftmost image or the distinct church entrance in the rightmost image. For the AFHQ dataset, SA can transfer the general color from the appearance image, but the resulting images strongly resemble the original structure image with minimal semantic changes. In contrast, our method preserves the general shape of the target structure while adapting its precise geometry to better

Table 1. Quantitative Comparison. We measure the level of structure preservation and appearance fidelity across all methods and various domains. To measure structure preservation, we calculate the mean IoU between binary masks extracted from the input structure image and the generated image. For appearance fidelity, we compute the distances between the Gram matrices of the input appearance image and the generated image.

Domain	Structure Preservation ↑			
	Swapping AE	SpliceViT	DiffuseIT	Ours
Buildings	0.82	0.56	0.79	0.76
Animal Faces	0.59	0.71	0.96	0.68
Animals	N/A	0.71	0.80	0.75
Cars	N/A	0.93	0.94	0.88
Birds	N/A	0.66	0.77	0.70
Cakes	N/A	0.64	0.66	0.61
Average	0.71	0.70	0.82	0.73
Domain	Appearance Fidelity ↓			
	Swapping AE	SpliceViT	DiffuseIT	Ours
Buildings	0.88	0.61	1.05	1.24
Animal Faces	0.58	1.11	1.74	0.27
Animals	N/A	2.53	2.89	1.41
Cars	N/A	0.56	0.50	0.21
Birds	N/A	0.14	0.15	0.41
Cakes	N/A	0.41	0.49	0.21
Average	0.73	0.89	1.14	0.62

capture the visual characteristics of the appearance image. For example, in the buildings domain, our method integrates the distinctive purple towers from the appearance image in the third column into the structure of the Eiffel Tower. Moreover, for the AFHQ dataset, our method adapts the target appearance and semantics to the shape of the cat, e.g., preserving the cat’s pointed ears.

In Figure 9 we present a comparison with techniques supporting appearance transfer between objects found in natural images. First, SpliceViT can transfer appearance between objects with similar shapes and viewpoints such as between the zebra and tiger in the first row or oranges and apples in the bottom row. However, when the two objects differ significantly in their visual characteristics, SpliceViT fails to find meaningful semantic correspondences. This results in heavy artifacts in the outputs, as seen in the second and third rows. Notably, SpliceViT requires a per-image generator tuning spanning dozens of minutes on a commercial GPU. While DiffuseIT attains results comparable to SpliceViT without the need for per-image model training, it still struggles to achieve high-quality transfer results in natural images. In contrast, our method can accurately transfer appearance between objects that vary in the number of instances (first and last rows) and between objects differing in shape (third and second-to-last row) and viewpoint (second and fourth rows). Moreover, our approach operates

Table 2. User Study. We asked respondents to select which set of images they most preferred based on their faithfulness to the input structure and appearance as well as the overall quality of the generated images. Results are averaged across all responses.

Buildings			
Method	Structure	Appearance	Overall Quality
Swapping AE	44.3%	3.1%	20.9%
SpliceViT	2%	17.7%	2.3%
DiffuseIT	16.4%	2.9%	10.4%
Ours	37.3%	76.3%	66.4%
Animals, Cars, Cakes, Birds			
Method	Structure	Appearance	Overall Quality
SpliceViT	11.0%	21.4%	9.6%
DiffuseIT	44.8%	8.5%	30.3%
Ours	44.2%	70.1%	60.1%

in a zero-shot setting while requiring no external models to guide the disentanglement process. Instead, we rely on the rich internal representations already captured by the model.

Quantitative Comparison We quantitatively evaluate each considered method in two aspects: (1) how well they preserve the source structure, and (2) how well the generated images depict the target appearance. To measure structure preservation, we first extract binary masks over the input structure images and corresponding output images using SAM [38]. We then measure the mean IoU of the output images with respect to the input structure images. As there is no standard automatic metric for assessing semantic-based appearance fidelity, we turn to the neural style transfer literature which has demonstrated that images with similar styles tend to have similar Gram matrices [20]. As such, we measure the L_2 distance between the Gram matrices of the input style and output images computed along five intermediate layers of a pretrained VGG19 [69] network.

As Swapping Autoencoder [56] is limited in its supported domains, we compute the above metrics across six domains (buildings, animal faces, animals, cars, birds, and cakes). For each domain, we selected 20 structure-appearance input pairs. Results are displayed in Table 1. As shown, our method demonstrates comparable performance to the alternative methods across all domains in both structure preservation and appearance fidelity. It is worth noting that achieving faithful semantic transfer often necessitates minor structure modifications. For example, merely transferring the general color scheme between images would lead to a high mean IoU, but would fail to capture the true semantics of the appearance image. Our method offers a favorable balance between preserving the precise input geometry and capturing the prominent semantics of the target appearance, as also supported by our qualitative evaluations.

User Study Finally, we conduct a user study to analyze all techniques across five object domains (buildings, animals, cars, cakes, and birds). For each domain, we selected multiple structure-appearance input pairs and generated transfer results using each of the four considered methods. Additional details on the evaluation setup are provided in Appendix A. For each pair, participants were tasked with evaluating the results based on three key aspects: (1) how well the source structure was preserved, (2) how well the output depicted the target appearance, and (3) the overall quality of the generated image. Participants were presented with the outputs from all relevant methods and were asked to select the most favorable result for each aspect.

Results are presented in Table 2 where the final score for each method is calculated by averaging the number of times participants selected that approach across all questions. In the buildings domain, Swapping Autoencoder outperforms all methods, which is likely due to its per-domain training. However, our method achieves a comparable level of structure preservation while significantly surpassing all other methods in the ability to capture the target appearance and generate high-quality images. In the remaining four domains, our method consistently outperforms both SpliceViT and DiffuseIT in appearance preservation and quality while achieving better or comparable structure preservation.

4.2. Ablation Study

Finally, we perform an ablation study to validate the key design choices of our method. Specifically, we assess the contribution of (1) the attention map contrasting operation, (2) the AdaIN normalization over the noised latent code, and (3) our appearance guidance technique performed over the noise estimates of the denoising network. The results are presented in Figure 10. For our baseline, we simply swapping the standard self-attention layer with our cross-image attention layer. As shown, while the general semantics are transferred from the appearance images to the structure images, many artifacts are present in the outputs. In each subsequent row, we add an additional component to our technique, with the final row representing our complete method. As shown, applying the contrasting operation significantly reduces the artifacts present in the baselines. By employing the AdaIN operation, we can better refine the general color distribution of the output as can be seen in the second column. Finally, incorporating the appearance guidance throughout the denoising process significantly improves the overall quality of results by refining finer-level details in the image. For example, observe the purple towers present in the leftmost column or the strawberries on the cake in the fourth column.

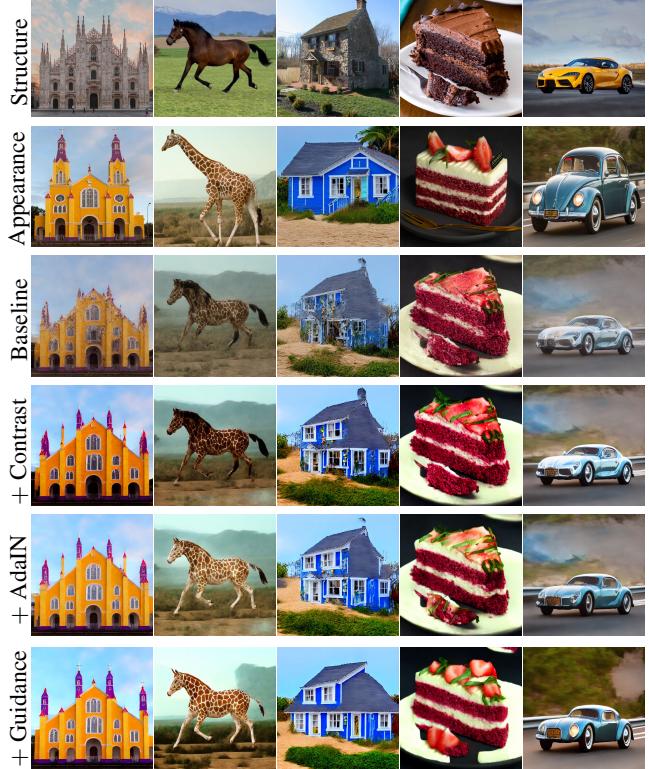


Figure 10. Ablation Study. In each row, we add an additional component of our appearance transfer scheme. Images in the bottom row represent results obtained by our complete method.

5. Limitations and Discussion

While we have demonstrated the effectiveness of our cross-image attention mechanism for zero-shot appearance transfer, several limitations should be considered. First, our method relies on the ability of the generative model to establish accurate correspondences between subjects in the two input images. As a result, transferring appearance between subjects in the images that do not share semantics (e.g., belong to different domains) can be more challenging, see the first two rows of Figure 11. Next, our method relies on inverting the input structure and appearance images into the latent space of the image diffusion model. In cases where the inversion fails to reconstruct the input or inverts the images into less editable latent codes, our transfer introduces unwanted artifacts. Specifically, the inversion method used in our approach may exhibit sensitivity to the random seed employed for the inversion, as evident in the bottom row of Figure 11 where the leg of the output may vary between random seeds. Achieving accurate, yet highly-editable inversions within the context of diffusion models remains an open problem. We believe that additional progress in this area will contribute to improved performance in downstream tasks such as appearance transfer.

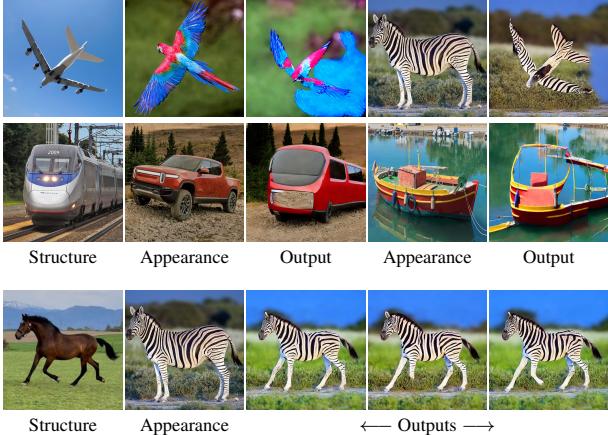


Figure 11. Limitations. Our method may struggle to transfer appearance between objects that do not share strong semantics (first two rows). Moreover, the quality of our transfer relies on the quality and editability of the inversion and may vary depending on the random seed used for the DDPM inversion (bottom row).

6. Conclusions

We have introduced a novel zero-shot approach that enables semantic-based appearance transfer between objects found in natural images. Importantly, our method demonstrates that this transfer is possible without requiring any model training or user-provided conditioning. Furthermore, this transfer can be achieved even when the objects vary in shape, size, or viewpoint. After examining the components of the self-attention layers — the queries, keys, and values — we introduced the Cross-Image Attention layer. This layer implicitly establishes semantic correspondences between objects by mixing the queries, keys, and values corresponding to two *different* images. We then introduced three extensions to reduce the domain gap caused by our mixing operation, accomplished through the manipulation of the noised latent codes and the internal representations of the denoising model. By leveraging the iterative denoising process, our method attains a *gradual* appearance transfer, encouraging the generation of more realistic, high-quality images.

We hope that our work encourages further exploration into the semantics of the internal representations within these powerful generative models. We believe that a deeper understanding of these representations can enable their utilization in addressing a diverse set of generative tasks with minimal user intervention while functioning in a zero-shot manner.

Acknowledgements

We would like to thank Michael Cohen and Lior Shapiro for their support and Amir Hertz, Dani Lischinski, Gal Metzer, Rinon Gal, and Yael Vinker for their insightful feedback. This work was funded by a research gift from Meta.

References

- [1] Kfir Aberman, Jing Liao, Mingyi Shi, Dani Lischinski, Baoquan Chen, and Daniel Cohen-Or. Neural best-buddies: Sparse cross-domain correspondence. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 3
- [2] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *arXiv preprint arXiv:2206.02779*, 2022. 3
- [3] Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, and Xi Yin. Spatext: Spatio-textual representation for controllable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18370–18380, 2023. 3
- [4] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers, 2023. 1
- [5] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023. 3
- [6] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7–13, 2006. Proceedings, Part I 9*, pages 404–417. Springer, 2006. 3
- [7] Saguy Benaim, Ron Mokady, Amit Bermano, and Lior Wolf. Structural analogy from a single image pair. In *Computer Graphics Forum*, pages 249–265. Wiley Online Library, 2021. 2
- [8] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 3
- [9] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. *arXiv preprint arXiv:2304.08465*, 2023. 2, 3
- [10] 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. 2
- [11] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 3
- [12] Seokju Cho, Sunghwan Hong, and Seungryong Kim. Cats++: Boosting cost aggregation with convolutions and transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3
- [13] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3
- [14] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains.

- In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020. 7
- [15] Tomer Cohen and Lior Wolf. Bidirectional one-shot unsupervised domain mapping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1784–1792, 2019. 2
- [16] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance, 2022. 3
- [17] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35:16890–16902, 2022. 1
- [18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2
- [19] Dave Epstein, Allan Jabri, Ben Poole, Alexei A. Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *arXiv preprint arXiv:2306.00986*, 2023. 1, 2, 3
- [20] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style, 2015. 8
- [21] Leon A Gatys, Alexander S Ecker, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. Controlling perceptual factors in neural style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3985–3993, 2017. 2
- [22] Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. Expressive text-to-image generation with rich text. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3
- [23] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 2, 3
- [24] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2
- [25] Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised semantic correspondence using stable diffusion, 2023. 3
- [26] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2, 3
- [27] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2, 5
- [28] Sunghwan Hong, Seokju Cho, Seungryong Kim, and Stephen Lin. Integrative feature and cost aggregation with transformers for dense correspondence. *arXiv preprint arXiv:2209.08742*, 2022. 3
- [29] Lianghua Huang, Di Chen, Yu Liu, Shen Yujun, Deli Zhao, and Zhou Jingren. Composer: Creative and controllable image synthesis with composable conditions. 2023. 3
- [30] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 2, 6
- [31] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpm noise space: Inversion and manipulations. *arXiv preprint arXiv:2304.06140*, 2023. 4, 14
- [32] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2
- [33] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015. 3
- [34] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. Neural style transfer: A review. *IEEE transactions on visualization and computer graphics*, 26(11):3365–3385, 2019. 2
- [35] Oren Katzir, Dani Lischinski, and Daniel Cohen-Or. Cross-domain cascaded deep translation. In *European Conference on Computer Vision*, 2020. 2
- [36] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Conference on Computer Vision and Pattern Recognition 2023*, 2023. 3
- [37] Anant Khandelwal. Infusion: Inject and attention fusion for multi concept zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3017–3026, 2023. 3
- [38] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 8
- [39] Nicholas Koltkin, Jason Salavon, and Gregory Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10051–10060, 2019. 2
- [40] Gihyun Kwon and Jong Chul Ye. Diffusion-based image translation using disentangled style and content representation. *arXiv preprint arXiv:2209.15264*, 2022. 2, 6
- [41] Yumeng Li, Margret Keuper, Dan Zhang, and Anna Khoreva. Divide & bind your attention for improved generative semantic nursing, 2023. 3
- [42] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 3
- [43] Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. Visual attribute transfer through deep image analogy. *arXiv preprint arXiv:1705.01088*, 2017. 2
- [44] Jun Hao Liew, Hanshu Yan, Daquan Zhou, and Jiashi Feng. Magicmix: Semantic mixing with diffusion models. *arXiv preprint arXiv:2210.16056*, 2022. 3

- [45] Jianxin Lin, Yingxue Pang, Yingce Xia, Zhibo Chen, and Jiebo Luo. Tuigan: Learning versatile image-to-image translation with two unpaired images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 18–35. Springer, 2020. 2
- [46] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30, 2017. 2
- [47] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 3
- [48] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *arXiv preprint arXiv:2305.14334*, 2023. 3
- [49] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 3
- [50] Juhong Min and Minsu Cho. Convolutional hough matching networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2940–2950, 2021. 3
- [51] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models, 2023. 1, 2
- [52] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1, 2, 3
- [53] Dolev Ofri-Amar, Michal Geyer, Yoni Kasten, and Tali Dekel. Neural congealing: Aligning images to a joint semantic atlas. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19403–19412, 2023. 3
- [54] 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. 2
- [55] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 319–345. Springer, 2020. 2
- [56] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei A. Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. In *Advances in Neural Information Processing Systems*, 2020. 1, 2, 6, 7, 8, 14
- [57] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. *arXiv preprint arXiv:2302.03027*, 2023. 3
- [58] Or Patashnik, Daniel Garabi, Idan Azuri, Hadar Averbuch-Elor, and Daniel Cohen-Or. Localizing object-level shape variations with text-to-image diffusion models, 2023. 2, 3, 6, 14
- [59] William Peebles, Jun-Yan Zhu, Richard Zhang, Antonio Torralba, Alexei A Efros, and Eli Shechtman. Gan-supervised dense visual alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13470–13481, 2022. 3
- [60] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1, 3
- [61] Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *arXiv preprint arXiv:2306.08877*, 2023. 3
- [62] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelović, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. *Advances in neural information processing systems*, 31, 2018. 3
- [63] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2, 3, 14
- [64] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. 3
- [65] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1, 3
- [66] Paul Hongsuck Seo, Jongmin Lee, Deunsol Jung, Bohyung Han, and Minsu Cho. Attentive semantic alignment with offset-aware correlation kernels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 349–364, 2018. 3
- [67] Arseniy Shakhmatov, Anton Razhigaev, Aleksandr Nikolichev, Vladimir Arkhipkin, Igor Pavlov, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky 2. <https://github.com/ai-forever/Kandinsky-2>, 2022. 1, 3
- [68] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net, 2023. 14
- [69] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 8
- [70] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Learning local feature descriptors using convex optimisation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8):1573–1585, 2014. 3
- [71] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, Yuan Hao, Irfan Essa, Michael

- Rubinstein, and Dilip Krishnan. Styledrop: Text-to-image generation in any style, 2023. 2
- [72] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 14
- [73] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *arXiv preprint arXiv:2306.03881*, 2023. 2, 3
- [74] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10748–10757, 2022. 1, 2, 6
- [75] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. pages 1921–1930, 2023. 2, 3
- [76] Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion models. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 3
- [77] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dual-gan: Unsupervised dual learning for image-to-image translation. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017. 2
- [78] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9036–9045, 2019. 2
- [79] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. 2023. 3
- [80] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 3
- [81] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1, 2

Appendix

A. Additional Details

Implementation Details We operate over Stable Diffusion v1.5 text-to-image model [63]. To invert the two input images, we apply the DDPM inversion technique introduced in [31] using their default hyperparameters and using the prompt “A photo of a *domain*” where *domain* denotes the domain of the object we wish to transfer (e.g., animal or building). For the denoising process, we employ the standard DDIM scheduler introduced by Song *et al.* [72] for 100 denoising steps.

For appearance transfer, we replace the conventional self-attention layers within the denoising network’s decoder at resolutions of 32×32 and 64×64 with our cross-image attention layers. However, we inject the keys and values only at a subset of the denoising timesteps. Specifically, for layers with a resolution of 32×32 , the injection is performed between timesteps 10 and 70, while for layers at a resolution of 64×64 the injection is applied between timesteps 10 and 90. At all other timesteps, our cross-image attention layer functions identically to the standard self-attention layer.

Additionally, we apply a contrast strength of $\beta = 1.67$ over the intermediate cross-image attention maps. For our appearance guidance, we set the guidance scale to $\alpha = 3.5$ and apply the AdaIN operation between the style and output noise latents between timesteps 20 and 100.

To compute the object masks used for the AdaIN operation, we use the unsupervised self-segmentation technique introduced in Patashnik *et al.* [58] using the domain name as the guiding noun. Finally, we apply the FreeU technique [68] over Stable Diffusion and find that doing so leads to fewer artifacts in the generated images.

Structure Injection Lastly, we explore a simple technique that we find helps to better preserve the original structure in I^{struct} for certain object domains. Instead of replacing the keys and values corresponding to z_t^{out} with those of z_t^{app} , we choose specific intervals where we replace K_{out} and V_{out} with the keys and values derived from z_t^{struct} . That is, the feature output at these timesteps is now defined as

$$\text{softmax} \left(\frac{Q_{out} \cdot K_{struct}^T}{\sqrt{d}} \right) \cdot V_{struct}. \quad (6)$$

We observe that this approach is effective for object categories containing finer-level structural details such as the ear of an animal. We find that performing this structure injection every five timesteps provides a favorable balance between faithfully transferring the target appearance to the output image while maintaining its original structure.

User Study Since Swapping Autoencoder [56] is limited to the buildings domain, we select eight structure-appearance pairs from the building domain and two pairs for the four other domains (animals, cars, cakes, and birds). This results in a total of 16 input pairs in total. Note that Swapping Autoencoder was not evaluated with respect to the four other domains as no trained models exist for these domains.



Figure 12. Additional appearance transfer results obtained by our method. For each set of images, we show transfer results between a single structure image (shown to the left) and three different appearance images (shown to the top).

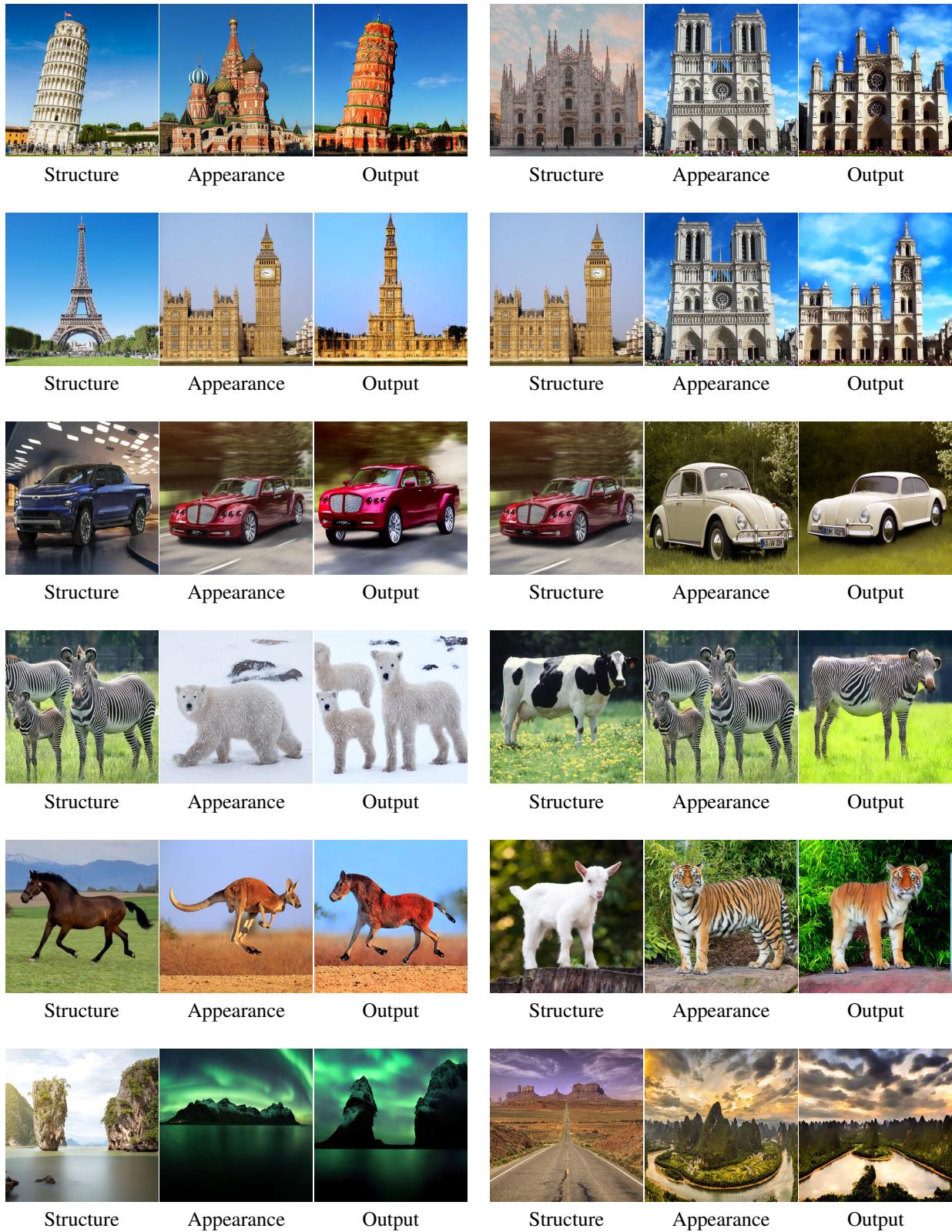


Figure 13. Additional appearance transfer results obtained by our method.



Figure 14. Additional appearance transfer results obtained by our method.

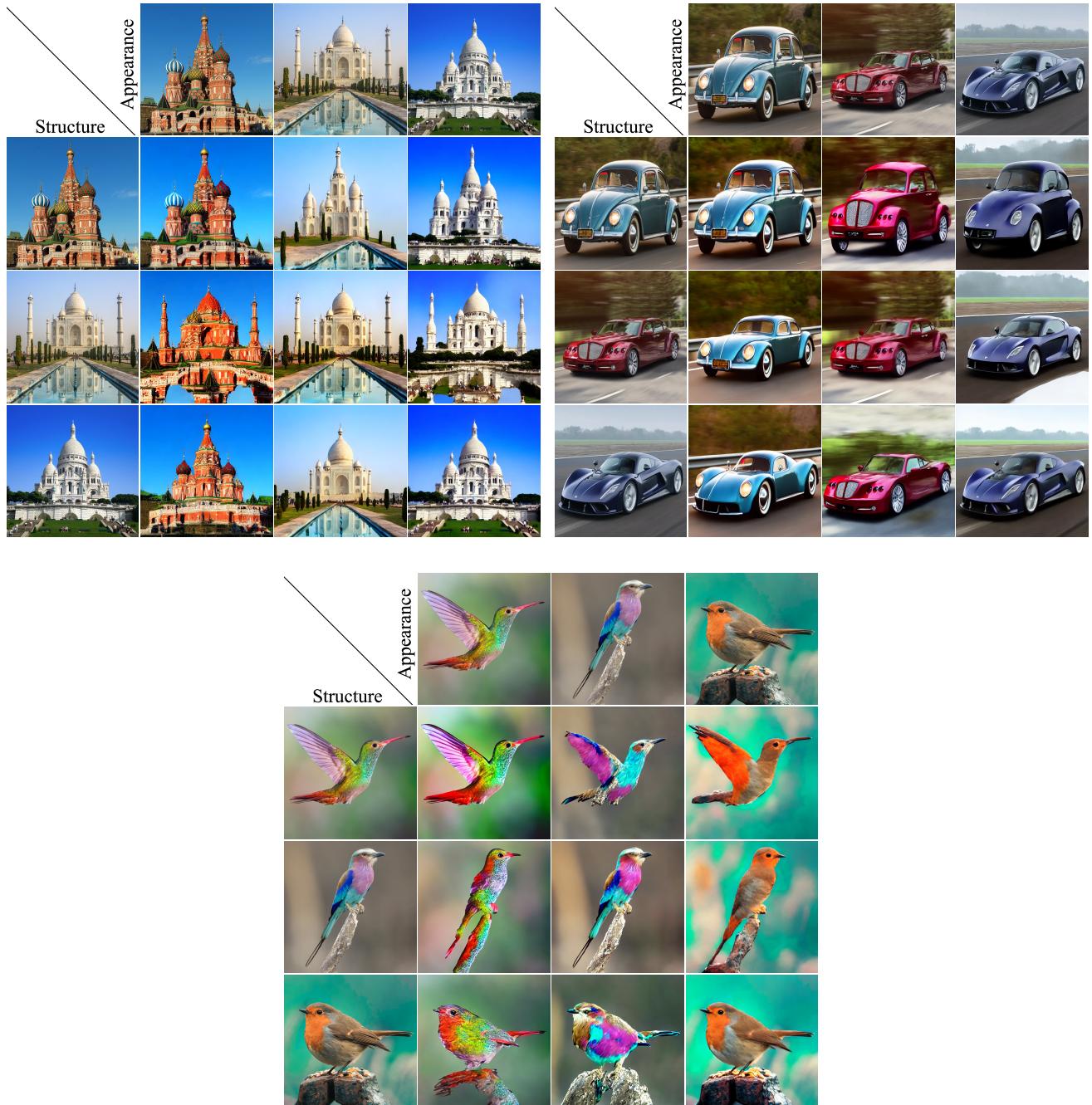


Figure 15. Enlarged versions of our appearance transfer results from Figure 6.