Diffusion Features in Stable Diffusion 3

Ilir Hajrullahu
LMU Munich
I.Hajrullahu@campus.lmu.de

Ludwig Degenhardt
LMU Munich
ludwig.degenhardt@lmu.de

Abstract

This practical project explores the adaptation of DIFT from Stable Diffusion 2 to Stable Diffusion 3 and evaluates its performance in semantic correspondence tasks. DIFT extracts image correspondences from diffusion model features, and we analyze how SD3's architectural shift from U-Net to transformer-based stacks affects its effectiveness. Through systematic experiments, we observe that SD3 performs worse, raising questions about transformer-based diffusion architectures for semantic correspondence tasks.

1. Introduction

The following is our project report for the practical "Image and Video Synthesis". Our topic, "Diffusion Features in recent T2I models," revolves around exploring semantic features in state-of-the-art text-to-image diffusion models. Our more concrete goal was to adapt the DIFT method for Stable Diffusion 3 and to investigate the adapted DIFT's performance in semantic correspondence tasks. This presents us with an interesting assignment, as the original version of DIFT had excellent results with Stable DIffusion 2-1, outperforming OPENCLIP and DINO in a PCK (Percentage of Correct Keypoints) benchmark test on the SPair-71k dataset [7]. Our task raises various questions: How does the new model Stable Diffusion 3 and its changed architecture influence DIFT's performance? Does it get better, stay the same, or even get worse? Can we learn something about diffusion models and their architectures in the process? Furthermore, adapting DIFT is interesting because of its possible applications, such as object recognition and edit propagation. Regarding the latter, DIFT can propagate edits across different images based on learned semantic correspondences, for example, adding a sticker to one image and propagating it accurately to another, even in different domains or categories [7].

The structure of this report is as follows: First we explain both DIFT and semantic correspondences, as well as the architectural changes between SD2 and SD3 in section 2. Section 3 presents how we proceeded over the course of our project and how we approached our experimental tasks, followed by section 4 demonstrating the results of these experiments. Section 5 then provides a discussion of our results and suggestions for future work, and lastly, section 6 concisely summarizes our project report.

2. Related work

2.1. DIFT and Semantic Correspondence

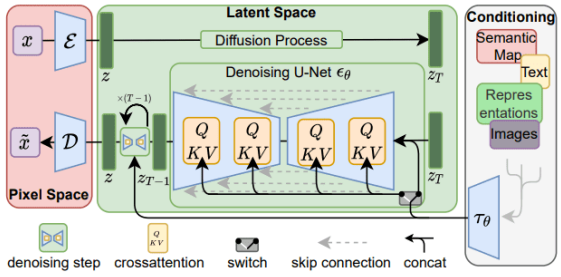
DIFT is a method introduced in the "Emergent Correspondence from Image Diffusion" paper, which explores how diffusion models inherently learn correspondences without explicit supervision [7]. Diffusion models, primarily designed for image generation, capture rich feature representations that we can leverage to establish correspondences between images. The name "DIFT" refers both to the method used to extract image features from pretrained diffusion models and to the extracted features themselves, which encode semantic, geometric, and temporal correspondences. In the original paper, the researchers implemented DIFT with two different open-source diffusion models, Stable Diffusion 2-1 [6], and Ablated Diffusion Model [1], with Stable Diffusion showing better results for correspondence extraction [7].

The DIFT method works by extracting correspondence-rich image features. It does so by simulating the forward diffusion process and passing noisy images through a pretrained diffusion model to obtain feature maps. These feature maps encode correspondences across images, and DIFT uses nearest-neighbor lookups using cosine similarity to find corresponding pixels.

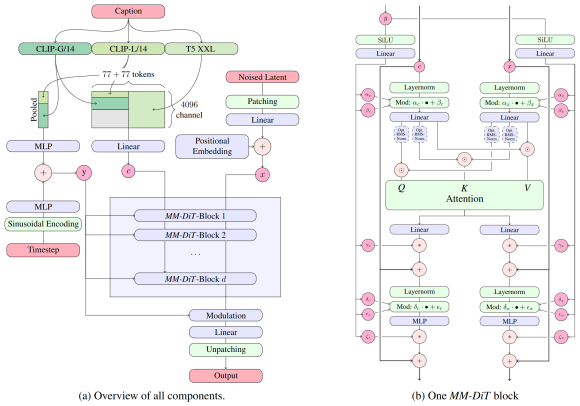
For our project, we focused on exploring semantic correspondences. Semantic correspondence refers to finding corresponding pixel locations across different images where objects share similar semantic meanings. This means that we can use DIFT to identify, for example, the ears of cats in different images, even if they are in different perspectives, contexts, or even styles, such as a photograph and a drawing.

2.2. Stable Diffusion 2 and 3

Stable Diffusion 3 (SD3) marks a departure from the traditional latent diffusion approach used in Stable Diffusion 2 (SD2). SD3 adopts MMDiT (Multimodal Diffusion Transformer), a more flexible architecture that processes multiple data modalities while enabling larger context windows. The idea of MMDiT is based on Diffusion Transformers that use separate weights for text and image and allow the flow of information between the two modalities [5]. In contrast to SD2's reliance on a single text encoder (often CLIP), SD3 incorporates three different text encoders to capture nuanced semantic information and accommodate more complex or lengthy prompts.



(a) SD2 Architecture as pictured in [6]



(b) SD3 Architecture as pictured in [2]

Figure 1. Stable Diffusion Architectures

Another prominent distinction is SD3's transition away from a U-Net-based noise prediction framework to a stack of diffusion transformers. This change aligns the model's denoising operations more closely with Transformer architectures, making cross-attention more adaptive when integrating textual features into the generation process. Moreover, SD3 employs rectified flow sampling, a refined technique for converting noisy intermediate representations into clear output images with fewer inference steps and improved coherence.

The underlying variational autoencoder (VAE) also receives a significant upgrade. Whereas SD2 typically used a 4-channel VAE, SD3 supports a 16-channel VAE, allowing a richer latent representation of the image space and providing finer-grained detail in generated images. Taken together, these enhancements illustrate SD3's push for larger capacities, deeper representations, and more efficient sampling, making it a noteworthy step forward in diffusion-based image generation.

3. Methodology

This work is based on adapting the DIFT codebase, aligned initially with SD2, to function seamlessly with the more recent SD3 model. Our overall aim was to maintain DIFT's ability to measure and visualize semantic correspondences between input prompts and generated images while taking advantage of the updated architecture and textual conditioning mechanisms introduced in SD3. Initially, we cloned the publicly available DIFT repository, which includes features to extract intermediate representations of the diffusion process and compare them across multiple images, prompts, and layers in the U-Net.

To integrate SD3, we made several key modifications to the original pipeline. One of the primary adjustments involved updating the text encoder interface, since SD3 often relies on larger, more sophisticated text-encoder backbones (such as T5). Our updated code handles the expanded prompt-token capacity and the different attention mechanisms these new encoders employ. Another significant change concerned the MMDiT structure. In some variants of SD3, the network is equipped with additional DiT-like blocks and an altered cross-attention mechanism. Consequently, we reworked how DIFT retrieves and processes these internal feature maps, which serve as the basis for measuring semantic alignment—effectively allowing us to see which parts of the image or latent representation the model deems most relevant for a given prompt.

In order to systematically evaluate these changes, we set up four distinct testing scenarios, all centered around the concept of semantic correspondence. The first scenario involved running DIFT on top of SD2.1 for baseline comparisons. By doing so, we wanted to get familiar with DIFT and get a baseline for future comparisons with SD3. We then moved on to the second scenario, where we tested DIFT with the newly integrated SD3 pipeline. This step aimed to verify how effectively DIFT captures semantic coherence with a bigger text encoder and transformer architecture and whether the new features in SD3 lead to improved or more detailed semantic maps.

The third testing scenario examined the effect of different latent-space configurations on DIFT's ability to measure correspondences. Models in both SD2 and SD3 typically apply a Variational Autoencoder (VAE) to compress images into fewer channels, and this step can

influence how well certain semantic signals are preserved. We, therefore, investigated 4-channel versus 16-channel autoencoders, exploring how differences in compression impact the detection of fine-grained features. For the final scenario, we directed DIFT to extract features from a specific transformer layer within the SD3 MMDiT blocks, allowing us to pinpoint at which stage the strongest semantic signals arise. This deeper look inside the network is especially relevant, as DIFT was initially designed with a focus on identifying how cross-attention layers in diffusion models correlate image and text.

All four tests were conducted on a set of 14 images chosen from the Spair-71k benchmark dataset to represent a variety of real-world and synthetic scenes, lighting conditions, and object categories [4]. These images enabled us to assess the consistency of DIFT's semantic alignment capabilities under different circumstances. By comparing how well the method performed in each scenario, we gained insight not only into the robustness of DIFT itself but also into the architectural and encoder-related changes in SD3. Overall, these experiments thoroughly investigated how deeper textual contexts, larger model capacity, and varying levels of latent compression affect the interpretation and visualization of semantic features within diffusion-based generation models.

4. Experiment results

Our experiments revealed a pronounced divergence in performance between SD2 and SD3, as shown in Figure 2. On the one hand, running DIFT with SD2 yielded crisp, well-defined semantic correspondences. As illustrated in the airplane example, the target image's heatmap clearly highlighted the areas of interest—such as the nose and especially the tail of the aircraft—corresponding to the cues in the source image. The heatmap uses a Viridis scale, with yellow demonstrating the highest semantic correspondence and purple the lowest, normalized to values between 0 and 1. The resulting visualizations demonstrated that most semantic information was captured in a concentrated and coherent manner, making it relatively easy to see how specific regions in the target image map back to salient features in the source.

By contrast, DIFT performance degraded substantially under SD3. The primary issue seemed to be that semantic information became diffusely "spread out" across the entire image rather than concentrated in a distinct region. This diffuse pattern was apparent across all 14 test images. In the case of the airplane, the target image heatmaps essentially displayed broad noise patterns rather than highlighting the plane's fuselage or tail section. Similarly, for the sheep example, neither the outline of the animal nor its features stood out in any localized area of the target's latent representation; instead, the visualization indicated a more uniform or random distribution of significance.

(a) Semantic Correspondence with SD2

(b) Semantic Correspondence with SD3

(c) Semantic Correspondence with SD2

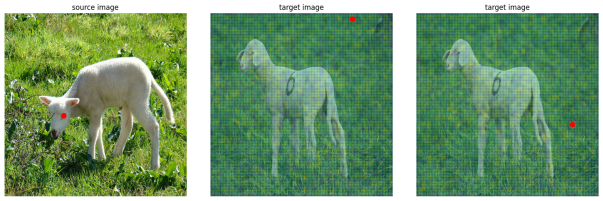
(d) Semantic Correspondence with SD3

Figure 2. Comparison for semantic correspondence of SD2 and SD3

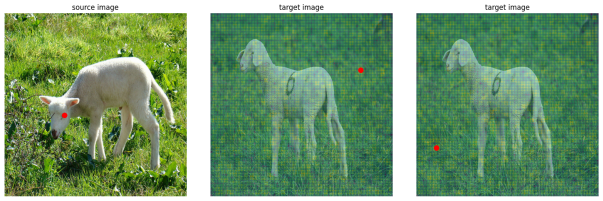
We also tested whether earlier layers of the SD3 model might contain cleaner, less "spread out" features. Specifically, we extracted and compared features at layer 2 and layer 6 out of the 12 available transformer blocks. Our initial hypothesis was that semantic structure might remain relatively intact at lower blocks before later transformer layers accumulate noise or disperse information.

Unfortunately, the results at both of these depths remained poor, as shown in Figure 3. Rather than isolating any coherent semantic points, the heatmaps still showed

generalized patterns across nearly the entire image with no distinct alignment to specific objects or regions of interest. The pattern was consistent across all 14 images we tested, suggesting that the problem is systematic rather than image-specific.



(a) Semantic correspondence of SD3 features after Transformer block 2



(b) Semantic correspondence of SD3 features after Transformer block 6

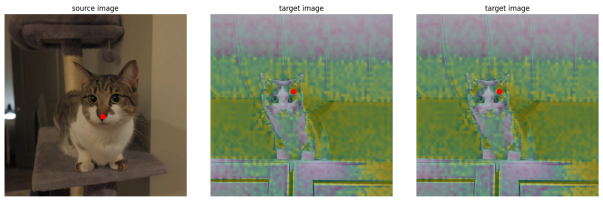
Figure 3. Semantic correspondence of SD3 after specific Transformer blocks

As for the results of our experiment using DIFT on vae-encoded images without utilizing Stable Diffusion, in both the 4-channel and 16-channel variants, DIFT could not extract meaningful semantic correspondences. However, we could notice the richer latent representation of the encoded image providing finer-grained details, making the target image's heatmap smoother, as shown in Figure 4.

Moreover, SD3 proved to be notably more resource-intensive during inference compared to SD2 without offering a corresponding improvement in semantic correspondence. GPU memory usage was higher, and inference times were longer, yet the DIFT method could not exploit any additional capacity that SD3 might have had to produce clearer correspondences. In effect, the newer diffusion architecture's improvements for general generation tasks did not translate into benefits for DIFT's targeted semantic alignment metrics.

5. Discussion and Future Work

When using DIFT with SD3, the semantic information we extract from the feature maps seems to be dispersed throughout the images, as visible in the according figures we provided in the section 4. To make the differences



(a) Semantic correspondence using 4-channel vae

(b) Semantic correspondence using 16-channel vae

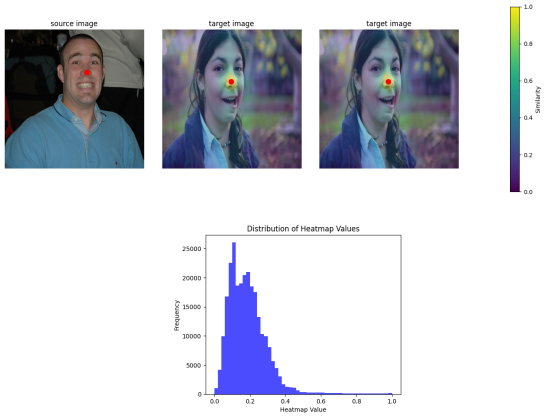
Figure 4. Semantic correspondence with vae-encoded images

between the heatmaps in the SD2 and SD3 versions of DIFT more comprehensible, we added a visualization of the heatmap's Viridis scale to the target images and plotted the distribution of semantic correspondence values, both visible in the example in Figure 5.

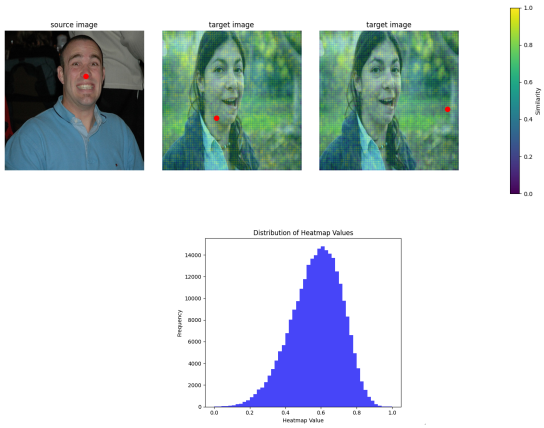
In the SD2 version of DIFT, most semantic correspondence values lie between 0.0 and 0.4. The higher values are located in semantically meaningful areas corresponding to the input area chosen in the source image. In the case of SD3, the heatmap's values align in a pattern resembling a standard distribution around medium-high values of 0.5 to 0.6, with only a few values around 0 and 1. This further supports our assumption that the semantic information is being spread through the image, with no indication of DIFT extracting meaningful semantic correspondences from the image features.

We assume that the reason for this behavior in the SD3 version of DIFT lies in its architecture. We hypothesize that when the transformer projects the input to a linear sequence to process it further, the semantic information is dispersed or mixed throughout the image.

We have some propositions for future work to explore the behavior of DIFT with state-of-the-art diffusion models, especially diffusion models with transformer architectures, more profoundly. First, it would be interesting to test the SD3-adapted DIFT on geometric and temporal correspondence tasks, which were outside the scope of our practical project. Another possibility would be to adapt DIFT for another different state-of-the-art text-to-image diffusion model, such as FLUX [3].



(a) Semantic correspondence values using SD2



(b) Semantic correspondence values using SD3

Figure 5. Visualization of semantic correspondence values of SD2 and SD3

By doing so, we could further explore if transformer-based architectures for text-to-image diffusion models perform poorly in semantic correspondence tasks or if there is a problem inherent to SD3. Exploring such possibilities could even come to influence architectural decisions for future models.

6. Conclusion

This project explored the adaptation of DIFT from Stable Diffusion 2 (SD2) to Stable Diffusion 3 (SD3), specifically evaluating its performance in semantic correspondence tasks. DIFT excelled in semantic correspondence tasks in its initial implementation using SD2. Our findings indicate a notable decline in DIFT's effectiveness when applied to SD3, raising questions about the suitability of transformer-based diffusion architectures for such tasks. The key observation was that semantic information in SD3's feature maps appeared more diffusely spread across images rather than being localized in meaningful regions, as seen with SD2. This degradation in correspondence quality may be attributed to SD3's architectural shift, particularly its use of transformer-based MMDiT instead of U-Net. Despite the model's improved generative capabilities, it does not translate into better performance in semantic correspondence tasks for DIFT-based feature extraction.

Future research should investigate whether this limitation is inherent to all transformer-based diffusion models or specific to SD3. Testing alternative architectures such as FLUX and evaluating how DIFT with SD3 performs in geometric and temporal correspondence tasks could provide further insights.

References

[1] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1

[2] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. 2

[3] Black Forest Labs. Flux ai official github repository: https://github.com/black-forest-labs/flux. 4

[4] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Spair-71k: A large-scale benchmark for semantic correspondence, 2019. 3

[5] William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. 2

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 1, 2

[7] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion, 2023. 1