Semantic-Guided Sketch-to-RGB Image Generation via Controlled Diffusion for Improved Sketch Recognition (Student Abstract)

Ritika Jain, Atul Kumar, Akshay Agarwal

Trustworthy BiometraVision Lab, IISER Bhopal, India
{ritika21, atulk23, akagarwal}@iiserb.ac.in

Abstract

Although deep networks excel on RGB images, their performance degrades sharply under severe domain shifts—such as sketch recognition, where color and texture cues are missing. In this work, we propose a novel pipeline that leverages semantic cues extracted from sketches to guide the synthesis of photorealistic RGB images using diffusion-based generative models. Our framework operates by extracting two crucial cues from the input sketch: semantic captions via the BLIP model and structural outlines via Canny edge detection. These cues are then integrated using ControlNet to guide a Stable Diffusion model, ensuring the synthesized RGB image is both semantically consistent with the content and structurally faithful to the original sketch. We evaluated our synthesized images by benchmarking classification performance. We trained standard architectures (from convolutional to transformer-based) on Tiny-ImageNet subsets and tested them on sketches, their synthesized counterparts, and the original RGB images. Experimental results demonstrate that our approach produces realistic, identity-preserving images, which significantly improve classification accuracy and effectively bridge the semantic gap. While BLIP-based captioning and ControlNet-guided diffusion are established methods, our contribution lies in their integration into a unified, caption-guided pipeline that enhances sketch-to-RGB translation with improved semantic consistency. The proposed method generalizes well across architectures, providing a scalable and cost-efficient solution for sketch-based image synthesis.

Introduction

Conventional image synthesis techniques that rely on GANs (Isola et al. 2017) and edge-guided inpainting models (Nazeri et al. 2019) often exhibit limited expressiveness or require paired datasets for effective training. Under sparse input conditions, style transfer methods (Gatys, Ecker, and Bethge 2016) fail to preserve structural semantics despite producing visually appealing results. Recent advances in diffusion models (Rombach et al. 2022) and their conditional variants (Zhang, Rao, and Agrawala 2023) allow for structure-preserving image generation from edge maps, while vision-language models like BLIP (Li et al. 2022) can extract semantic meaning by generating image captions from sparse sketches.

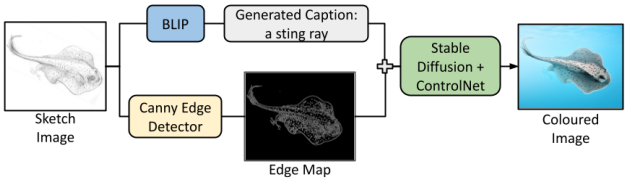

Figure 1: Proposed pipeline using BLIP to generate image captions, and ControlNet v1.1p (Canny version) with Stable Diffusion v1.5 for scribble/edge-guided diffusion-based sketch-to-RGB image conversion.

In our prior work (Jain, Agarwal, and Kumar 2025), we have explored domain shift in sketch-based recognition and analyzed how colorization or semantic augmentation can enhance cross-domain recognition. Building upon these insights, we present a pipeline that automatically generates textual captions from sketches, preprocesses edge information via Canny detection, and feeds them into a controlled diffusion model for RGB image synthesis. Unlike prior approaches, our method requires no paired sketch-RGB dataset, making it more scalable and adaptable across object categories. We evaluate the synthesized images by measuring their utility in downstream classification tasks. Our contributions are: (i) A novel caption-guided sketch-to-RGB generation pipeline using BLIP and ControlNet. (ii) A scalable approach that eliminates the need for paired datasets. (iii) A comprehensive evaluation across architectures shows improvements in visual fidelity, classification performance, and model generalization.

Proposed Approach for Sketch-to-RGB Generation

This work proposes a novel sketch-to-image synthesis pipeline integrating BLIP v1 for caption generation, Canny edge detection for spatial constraints, and Stable Diffusion v1.5 with ControlNet v1.1p (Canny) for image generation. Given a sketch image, as seen in Figure 1, the BLIP model generates a descriptive caption by maximizing the conditional likelihood of text given the visual input, thereby providing high-level semantic context. Here, we add a fixed prompt "realistic high quality coloured photograph of the

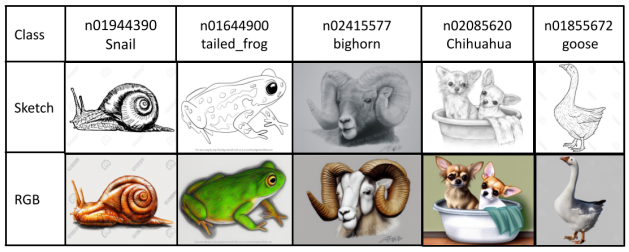

Figure 2: ImageNet sketches and their corresponding colored versions showcase how synthetic color integration changes the visual perception of images.

Model	RGB		Raw Sketch		Proposed Sketch	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
AlexNet	69.96	91.56	22.36	44.72	**27.24**$_{4.88\uparrow}$	**52.12**$_{7.40\uparrow}$
VGG19	74.64	92.36	35.40	57.04	**43.00**$_{7.60\uparrow}$	**63.88**$_{6.84\uparrow}$
ResNet50	70.20	89.88	12.44	29.52	**26.16**$_{13.72\uparrow}$	**48.04**$_{18.52\uparrow}$
EfficientNet-B7	59.80	83.56	16.16	33.40	**32.24**$_{16.08\uparrow}$	**52.72**$_{19.32\uparrow}$
ViT-B16	89.16	97.32	53.04	68.64	**59.56**$_{6.52\uparrow}$	**75.12**$_{6.48\uparrow}$

Table 1: Colorization (test-time adaptation) improves classification performance on sketch images, with synthetic colorization notably boosting performance on advanced models. Values underscored ↑ indicate the improvement in performance achieved using the proposed approach.

object on a plain white background, add natural colors, 4k resolution" at the end of the generated caption for more guidance. Simultaneously, the sketch is processed through the Canny Edge Detector to extract prominent edges, which are converted into a 3-channel scribble map to guide the spatial structure of the generated image. With a fixed maximum image size of 1024, these inputs are fed into the diffusion model, where a noising process is iteratively reversed using the caption embedding and edge features, with a guidance scale of 12 that controls prompt adherence. The denoising steps are optimized using the UniPCMultistepScheduler, resulting in a realistic, high-resolution RGB image that accurately aligns with the original sketch's content and contours. This pipeline effectively leverages pretrained models and classical image processing techniques to achieve controlled and coherent image synthesis from sparse inputs.

Experimental Setup

This research aims to generalize RGB-trained models to colorized sketches, establishing a baseline for test-time adaptation methods on unseen domains. To achieve this, we first generate a colorized RGB version of the raw sketches for 50 class samples (i.e., 50 images per class) using our proposed method. Subsequently, we fine-tune the final fully connected layer of five different architectures - AlexNet, ResNet50, VGG19, EfficientNetB7, and ViTB16 on the 50 classes from the Tiny ImageNet 200 training set (25,000 images). The models are then evaluated on the 50-class subset of the Tiny ImageNet 200 validation set, ImageNet sketch, and its corresponding generated RGB version. For training, we employ the ADAM optimizer for simpler models (AlexNet, ResNet50, and VGG19) and AdamW for more complex ones (EfficientNet-B7 and ViT-B16), with learning rates of 0.001 for all models except ViT, which uses a learning rate of 3e-5. All models are fine-tuned for 10 epochs, using a step-size scheduler for simpler models and cosine annealing for complex ones.

Experimental Results and Analysis

We first analyze our caption-guided colorization pipeline qualitatively. Figure 2 shows synthesized RGB versions of ImageNet sketches that successfully preserve structural contours while incorporating realistic textures and colors, resulting in visually rich representations. These qualitative results highlight the potential of the proposed pipeline to generate visually coherent and semantically meaningful images suitable for downstream tasks. Further, Table 1 reports Top-1 and Top-5 classification accuracies across multiple architectures on ImageNet-200 (RGB), ImageNet-Sketch (raw/original), and the proposed synthesized colorized sketches. All models experience a substantial drop in performance on sketches due to the absence of texture and color cues; for instance, ViT-B16 drops from 89.16% to 53.04% Top-1 accuracy, while ResNet50 and EfficientNet-B7 experience even larger reductions. After applying our caption-guided colorization, classification improves across all architectures, with ViT-B16 achieving 59.56% Top-1 accuracy and EfficientNet-B7 improving from 16.16% to 32.24%. CNN-based models, such as AlexNet and VGG19, also benefit, although to a lesser extent, indicating that deeper and transformer-based architectures derive more benefit from the additional semantic and color information. These results show that the synthesized images effectively bridge the domain gap between sketches and RGB images, enhancing model generalization under domain shift. Both qualitative and quantitative evaluations confirm our pipeline produces high-fidelity images by preserving structural integrity while introducing realistic colorization, thereby improving downstream classification performance. Despite this, we observed limitations when handling images with multiple objects, low contrast between the background and objects, or partially visible objects. These cases introduce ambiguity in control signal guidance, which we plan to address in future work.

Conclusion

We propose a cost-effective, semantic-guided framework for sketch-to-RGB image generation that integrates BLIP, ControlNet, and Stable Diffusion. This pipeline synthesizes realistic RGB images without requiring paired sketch-image datasets. Our results demonstrate that leveraging pretrained models for this image-to-image translation task improves classification performance under domain shifts. Evaluations across diverse architectures confirm the efficacy of the generated images in downstream classification tasks, establishing the approach as a scalable and generalizable solution for sketch-based image generation. Future work will focus on scaling this method from 50 to 1000 classes and exploring adaptive prompting strategies to enhance its applicability in real-world, sketch-based applications.

References

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image style transfer using convolutional neural networks. In *IEEE CVPR*, 2414–2423.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *IEEE CVPR*, 1125–1134.

Jain, R.; Agarwal, A.; and Kumar, A. 2025. Deep Models Under Domain Shift: A Sketch-Based Study. In *Women in Machine Learning Workshop@ NeurIPS 2025*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 12888–12900.

Nazeri, K.; Ng, E.; Joseph, T.; Qureshi, F. Z.; and Ebrahimi, M. 2019. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF CVPR*, 10684–10695.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF ICCV*, 3836–3847.