Tom and Jerry in Real Life - Translating Cartoon to Natural Images using Stable Diffusion

Sachin Salim (sachinks) Shrikant Arvavasu (ashri) Nowrin Mohamed (nowrin)

Abstract

In this work, we propose an unpaired image-to-image translation approach by harnessing the potential of Stable Diffusion and leveraging BLIP-based content transfer. Addressing the challenge of transforming cartoon scenes into hyper-realistic landscapes, our methodology taps into the raw power of Stable Diffusion, renowned for its dominance in generative modeling with limited paired data. This framework expertly conditions the generative process, facilitating seamless content transfer across domains, even in the absence of explicit correspondences. To enhance our model's interpretative prowess, we strategically integrate BLIP, a pre-trained image captioning juggernaut, effectively bridging the semantic gap between whimsical cartoons and the subtleties of natural imagery. The fusion of these cutting-edge methodologies not only yields visually compelling translations but also pushes the boundaries of what's achievable in unpaired image translation.

1. Introduction

Image-to-image translation, integral to generative modeling, finds applications across computer vision and augmented reality. Traditionally, models leverage paired datasets, where each source domain image corresponds to a target domain counterpart. However, the practicality of acquiring such paired training data is often limited. Unpaired image-to-image translation becomes pivotal in scenarios where establishing direct correspondences between source and target domain images poses challenges. Models in this context decode and transfer relevant attributes from the source to the target domain, preserving essential content and context even in the absence of explicit correspondences [21].

In the landscape of image synthesis, the effectiveness of stochastic generative modeling shines, generating diverse images within specific domains without requiring domain-specific knowledge. Recent strides in generative modeling focus on Deep Neural Networks (DNNs), particularly deep generative models (DGNNs), exemplified by Generative Adversarial Networks (GANs) [9], autoregressive models [11], flow-based models like NICE [6], RealNVP [7], and Glow [17], Variational Autoencoders (VAEs) [27], and Image Transformer [26]. Iterative generative models, including Denoising Diffusion Probabilistic Models (DDPM) [12] and Noise Conditional Score Networks (NCSN) [32], mark significant progress in generative modeling.

Our emphasis on translating cartoon images to realistic scenes highlights the effectiveness of unpaired translation in tackling challenges posed by exaggerated cartoon elements. The model adeptly discerns core content, balancing low-level features and high-level semantics for realistic adaptation. To address unpaired translation challenges, we leverage the Stable Diffusion framework, excelling in generative modeling for scenarios with limited paired data. Notably, Stable Diffusion effectively conditions the generative process without requiring direct correspondences, facilitating seamless content transfer across domains. Our approach integrates BLIP, a pre-trained image captioning model, to bridge the semantic gap between source and target domains. By harnessing BLIP's advanced captioning features, the model interprets and encodes visual information textually.

2. Related Work

2.1. Image-to-Image Translation

Typically, I2I translation is categorized into two main approaches: paired and unpaired methodologies.

2.1.1 Paired Image-to-Image Translation

Supervised methods learn input-output relationships using aligned image pairs. Early techniques employed pre-trained CNNs and Gram matrices [8] to separate content and style, preserving semantic details while allowing style variation. Recent approaches integrate GANs [10], where the generator produces data from random values, and the discriminator distinguishes real from generated data. Pix2Pix [15] provides a generalized adversarial framework using U-Net [29] to share information. BicycleGAN [35] enhances image reconstruction by combining CVAE-GAN [2] for latent code recovery.

2.1.2 Unpaired Image-to-Image Translation

Paired image-to-image (I2I) translation relies on aligned image pairs from both source and target domains, while unpaired approaches, such as CycleGAN [34], do not require such pairings. CycleGAN employs a Generative Adversarial Network (GAN) to transform source images $x_s \in X_s$ to target images $x_t \in X_t$, learning direct and reciprocal transformation paths $G_t(x_s)$, $G_s(x_t)$. The introduction of cycle-consistency loss $L_{cyc}(G_s, G_t)$ enforces coherence between real and generated domain images. Unsupervised Image-to-Image Translation Networks (UNIT) [22] assume a shared-latent space. Addressing multimodality, methods like Multimodal UNIT (MUNIT) [14] and Diverse Image-to-Image Translation via Disentangled Representations (DRIT++) [19] adopt disentangled representations to enable diverse translations from unpaired samples. In unpaired I2I, methods typically fall into two categories: two-side (e.g., CycleGAN [34], DualGAN [33]) enforcing cycle-consistency and one-side (e.g., DistanceGAN, GCGAN) preserving content via geometry distance or consistency measures between input and output. Recent advancements like U-GAT-IT [16] introduce attention mechanisms, while CUT [25] emphasizes maximizing mutual information through contrastive learning for improved translations.

2.2. Diffusion Models

Denoising Diffusion Probabilistic Models (DDPM) [12] introduce noise progressively to images, functioning effectively as a generative model by learning to reverse this corruption. The latent information derived from DDPM serves to interlink different image domains, establishing a connection between their respective latent spaces.

Diffusion Probabilistic Models [31] excel in density estimation and sample quality [5], employing UNet-like architectures tailored for image data biases. Optimizing their synthesis involves a reweighted objective [12] that balances image quality and compression. However, evaluating these models in pixel space poses challenges, including slow inference and high training costs. Strategies such as advanced sampling [18] and hierarchies [13] address some issues, though training on high-resolution images remains computationally intensive, necessitating costly gradient computations.

3. Methodology

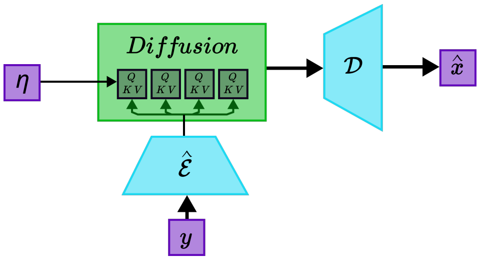

Figure 1. Model Architecture for Diffusion-based Image-to-Image Translation

In the realm of generative modeling, image-to-image translation can be solved using a conditional generation scheme. These models are trained to generate images in a specific target domain while maintaining relevant content from the source domain. The unique aspect of these models lies in their ability to assimilate and transfer relevant attributes from the source image to the target domain while preserving the core content and context.

In our project, we tackled the challenge of unpaired image-to-image translation using the Stable Diffusion framework, ideal for scenarios lacking readily available paired training data. Our aim was to transform cartoon images into natural scenes aligned with the cartoon's content. To achieve efficient content transfer, we leveraged BLIP (Bootstrapped Language Image Pretraining), a pre-trained image captioning model known for its advanced features in extracting and understanding visual content. By integrating BLIP features, our framework effectively interpreted and encoded visual content from cartoon images into textual representations. Subsequent sections provide a detailed breakdown of each framework component.

3.1. Stable Diffusion

Stable Diffusion is a high-quality generative modeling framework adapted from Latent Diffusion Models (LDMs) [28]. Diffusion models, in general, are Markovian generative models where, $\eta \sim \mathcal{N}(0, I)$ is the initial noise from which, a sample is generated using a reverse diffusion process. LDMs introduced a robust conditioning mechanism where a condition embedding $c = \varphi(C)$, where C is the raw conditioning input (such as an image, class label, or some text) and the model generates a sample $\hat{z}_\theta(\eta, c)$, which is then decoded using the latent decoder ($\mathcal{D}$) to get $\hat{x}_\theta(\eta, c)$. The model is trained to learn the noise added to get the latent code $z_t := \alpha_t z + \sigma_t \eta$ as follows,

$$z = \mathcal{E}(x) \tag{1}$$

$$\hat{\theta} := \min_{\theta} \ \mathbb{E}_{z,c,\eta,t} \left[\ \|\hat{z}_\theta(z_t, t, c) - z\|^2 \ \right] \tag{2}$$

Firstly, we trained a class-conditional diffusion model to perform generation in the target domain ($\mathcal{X}$) for the categories cat-only, mouse-only, and cat-and-mouse. We then freeze the diffusion model and make use of the cross-attention layers in the LDM to learn correspondences between the source domain ($\mathcal{Y}$) and the target domain $\mathcal{X}$ using a latent encoder. LDMs allow for source-conditioned generation by learning a mapping between their respective latent spaces using the cross-attention layers, where the query is learned from the source embedding and the key-value pair is learned from the target embedding as given in Equation 6. The diffusion model then accepts the projected target embedding and thus conditions the reverse diffusion process to generate images that align with the target domain. This formulation allows the diffusion model to learn content from the source image while maintaining the style of the target domain.

$$(x, y) \sim (\mathcal{X},\ \mathcal{Y}) \tag{3}$$

$$z \sim \mathcal{E}(x),\ y \sim \varphi(y) \tag{4}$$

$$Q = W_Q \cdot y,\ K = W_K \cdot z; V = W_V \cdot z \tag{5}$$

$$z_t^C = \text{Attention}(Q, K, V) = \ \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V \tag{6}$$

$(\cdot)$ refers to flattening and learnable-matrix multiplication.

An additional challenge in translating Tom and Jerry images to natural scenes is teaching the model to associate Tom with a cat and Jerry with a mouse. To address this, we introduce class conditioning using dataset class partitions (refer to Section 4.1). Class embeddings are created for each category, and the Latent Diffusion Model (LDM) is conditioned to generate images corresponding to the class label of the source image.

3.2. BLIP: Image Captioning

BLIP is a multi-modal model that is trained using a large web corpus to learn contextual information from images and corresponding text pairs [20]. Firstly, we finetune a BLIP model to learn to generate scene descriptive captions for both the source and target domain. Once trained, we hypothesize that we should observe a high degree of similarity between the captions for the source image and the target image. We thus propose a regularization using the fine-tuned BLIP encoder (ε) as given in Equation (7) that can be used to train the diffusion model to transfer the content of the source image to the target image.

$$\mathcal{L}_{reg,t} = \|\varepsilon_{\text{BLIP}}(y) - \varepsilon_{\text{BLIP}}(\mathcal{D}(z_t))\|^2 \tag{7}$$

Combining all of these modules, we implemented the following workflow in our project.

1. Pre-train a class-conditional latent diffusion model ($\hat{\theta}$, $\mathcal{E}, \mathcal{D}$) to generate natural images for cat-only, mouse-only, and cat-and-mouse cases.
2. We finetune a BLIP model to generate descriptive captions for both real cat and mouse images and Tom and Jerry images.
3. We freeze the diffusion model (except the cross-attention layers) and train an encoder (φ) to encode the Tom and Jerry images and generate natural images for the input Tom and Jerry images as shown Figure 1. During fine-tuning, the model would be trained using only the regularization loss given in Equation (7).

4. Experiments and Results

4.1. Image Datasets and Captioning

Datasets for the projects were painstakingly obtained from varied sources (see Appendix 6.1) and organized as follows. After dataset pruning (Appendix 6.1), there was a total of 5000 images in each source/cartoon and target/natural domains. They were further divided into three classes: those with only subject A (Tom/cat), with only subject B (Jerry/mouse), and those with both subjects A and B. This data was then uniformly separated into training and test datasets in the standard 80-20 ratio.

From the training set, 1200 images were randomly picked and the captions for the image actions were generated using GPT-4's [24] paid APIs [23] available. This image-caption pair data was leveraged to fine-tune BLIP, enhancing its performance for our specific use case through image-action caption alignment.

4.2. Implementation and Experimentation

Our project leveraged Python 3.9 and the PyTorch framework for model initialization and training. We utilized the open-source PyTorch implementation provided by the Computer Vision and Learning Group at LMU Munich [28] for our diffusion models and we used the HuggingFace framework for the BLIP model. The training process was executed on an NVIDIA A40 GPU cluster, equipped with 48 GB of RAM per GPU, to ensure efficient computation and data handling. The training duration was carefully managed across different model components:

- The latent diffusion model was trained for 20 hours, encompassing both the latent space and the diffusion model itself.
- The source-conditioned generation model was trained for a focused duration of 8 hours.
- The BLIP model was trained for 15 hours to fine-tune its performance for highly descriptive captions.

For data management, our dataset was methodically divided, allocating 80% for training and 20% for testing purposes. This split ensured a robust training process while reserving a significant portion of data for model evaluation. The training was conducted with an optimized batch size of 4 to balance the computational load and learning efficiency. For the image captioning model, we fine-tuned the BLIP model to get a BLEU (BiLingual Evaluation Understudy) score of 0.83 when compared against the gathered captions.

During training, we found the number of attention heads in the cross-attention layers of the diffusion model to be a

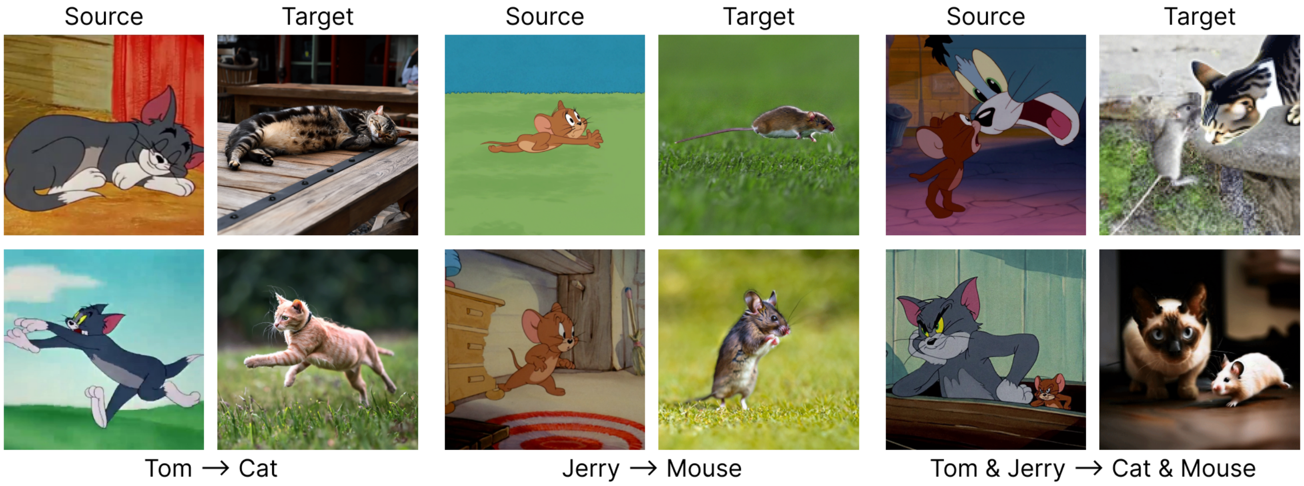

Figure 2. Comparison of Translation Results - Two optimal outcomes for each translation class: 'Tom to Cat,' 'Jerry to Mouse,' and 'Tom & Jerry to Cat & Mouse.' Note the fidelity variation; combined subjects exhibit comparatively lower faithfulness in the generated images.

deciding factor for the model performance. Thus, we experimented with 8, 16, 32, and 64 attention heads in each of the attention layers. The performance of the model in each of these settings is illustrated in Table (1).

Input Type	Attention Heads	FID ↓	MS-SSIM ↑
Tom-only	8	66.74	0.29
	16	58.53	0.34
	32	49.59	0.39
	64	46.32	0.43
Jerry-only	8	70.31	0.25
	16	67.54	0.32
	32	63.76	0.37
	64	58.33	0.38
Tom and Jerry	8	84.11	0.21
	16	86.13	0.24
	32	78.53	0.27
	64	75.21	0.28

Table 1. Experimental Results: Model performance on varying the number of attention heads of the cross-attention layers in the diffusion model.

4.3. Results

Table (2) presents a detailed analysis of our image translation models, showcasing the mean FID (Fréchet Inception Distance) scores and MS-SSIM (Multi-Scale Structural Similarity Index Measure) scores. These metrics were calculated for the three translation scenarios: Tom-to-cat, Jerry-to-mouse, and Tom-and-Jerry to cat-and-mouse. In addition to the quantitative analysis, Figure 2 provides visual evidence of the models' translation capabilities. This illustration includes selected examples of successful translations. Notably, the figure highlights a key observation: single-subject images (such as Tom-only or Jerry-only) yielded more accurate translations compared to multi-subject images. This outcome is reflective of the underlying data distribution utilized in our training dataset, suggesting a higher model proficiency with simpler, single-subject images. We also found that the results were sub standard in few cases for each class which is recorded and show in Figure 3 in the Appendix section.

Input Type	FID ↓	MS-SSIM ↑
Tom-only	46.32 ± 1.34	0.43 ± 0.09
Jerry-only	58.33 ± 2.25	0.38 ± 0.12
Tom and Jerry	75.21 ± 3.73	0.29 ± 0.14

Table 2. An overview of the results of the diffusion models on generating natural images for Tom-only, Jerry-only, and Tom-and-Jerry images. On average, the model performs better when trained on image pairs with only one subject, since their samples are abundant.

5. Conclusions and Discussions

This project proposes Stable Diffusion, aiming for faithful unpaired cartoon-to-natural image translation by leveraging BLIP to guide conditional target generation. Maintaining content fidelity faced hurdles due to cartoon subjects often depicted in unnatural actions or severe deformations. Future endeavors could explore learning segmentation maps to better preserve action postures during translation.

Additionally, the inference time for image generation with LDM was suboptimal. To address this, potential enhancements involve modifying the Markovian process, mirroring strategies found in denoising diffusion implicit models. This alteration seeks to optimize and expedite the image generation process within the framework, potentially alleviating the computational burden associated with slow inference in LDM.

References

[1] Niek Andresen, Manuel Wollhaf, Katharina Hohlbaum, Lars Lewejohann, Olaf Hellwich, Christa Thone-Reineke, and Vitaly Belik. Towards a fully automated surveillance of well-being status in laboratory mice using deep learning: Starting with facial expression analysis. *Plos one*, 15(4):e0228059, 2020. 6

[2] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Cvae-gan: fine-grained image generation through asymmetric training. In *Proceedings of the IEEE international conference on computer vision*, pages 2745–2754, 2017. 1

[3] Baskar, Bala. Tom and Jerry Image classification. https://www.kaggle.com/datasets/balabaskar/tom-and-jerry-image-classification/, 2021. Kaggle Dataset. 6

[4] Crawford, Chris. Cat Dataset. https://www.kaggle.com/datasets/crawford/cat-dataset/data, 2017. Kaggle Dataset. 6

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[6] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation, 2015. 1

[7] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp, 2017. 1

[8] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015. 1

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2014. 1

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1

[11] Karol Gregor, Ivo Danihelka, Andriy Mnih, Charles Blundell, and Daan Wierstra. Deep autoregressive networks, 2014. 1

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 1, 2

[13] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23(1):2249–2281, 2022. 2

[14] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018. 2

[15] 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. 1

[16] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwanghee Lee. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *arXiv preprint arXiv:1907.10830*, 2019. 2

[17] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2018. 1

[18] Zhifeng Kong and Wei Ping. On fast sampling of diffusion probabilistic models. *arXiv preprint arXiv:2106.00132*, 2021. 2

[19] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018. 2

[20] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 3

[21] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 1

[22] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30, 2017. 2

[23] OpenAI. GPT-4 Vision Preview APIs. https://platform.openai.com/docs/api-reference/chat/create, 2023. 3

[24] OpenAI. Gpt-4 technical report, 2023. 3

[25] 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

[26] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer, 2018. 1

[27] Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. Variational autoencoder for deep learning of images, labels and captions, 2016. 1

[28] 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 (CVPR)*, pages 10684–10695, 2022. 2, 3

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 1

[30] Michal Shaharabani. cat-mouse dataset. https://universe.roboflow.com/michal-shaharabani/cat-mouse, 2022. visited on 2023-12-02. 6

[31] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 2

[32] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 1

[33] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017. 2

[34] 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. 2

[35] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. *Advances in neural information processing systems*, 30, 2017. 1

6. Appendix

6.1. Datasets

These datasets were obtained from multiple different sources. The cat dataset was obtained from Kaggle [4]. There were over 9,000 images of cats with annotated facial features. The annotation data was dropped from our analysis. Most of the images were dropped to pick the best 800 images. The mice data was obtained from [1]. It included 18273 images with a mean size of 857 x 879px. We picked the best and varied 800 images from it. From [30], we collected data with both cats and mice and picked 400 images.

The Tom and Jerry data was collected from Kaggle [3]. This dataset contained more than 5k images (exactly 5478 images) extracted from some of Tom & Jerry's show videos, that are available online. The images were already separated into different folders: the ones containing only 'Tom', only 'Jerry', and the folder containing both.

Once the data was collected, we removed those images where the subject was either too small, significantly obscured or distorted. After this step, we made sure to keep 800, 800, and 400 images respectively in each of the three classes (subjects A, B, and both). All images underwent a standardization process where they were uniformly cropped to a size of 512x512 pixels.

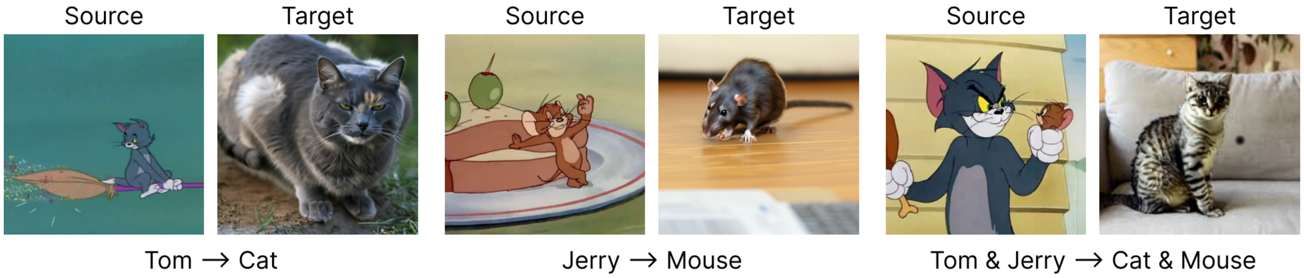

Figure 3. Comparison of Translation Results - A sub optimal outcome for each translation class: ‘Tom to Cat,’ ‘Jerry to Mouse,’ and ‘Tom & Jerry to Cat & Mouse.’