Literature Review

CSE499A – Section 15 **Group 4**

Name	Student ID	Reference Reviewed
Md. Minhajul Islam	2211022042	Reference 1, 2, 3
Nur Ibne Kawsar Zitu	2212021042	
Kazi Tazrian Mon	2132798642	Reference 4,5,6

Biswas et al. proposed a new method called PIXELS, which allows users to edit images using example pictures, enabling flexible modification of individual pixels or regions. The approach is based on Latent Diffusion Models, operating on lower-dimensional representations, and uses encoders and decoders to map inputs and outputs across pixel-latent manifolds. They also introduced an algorithm for progressive image editing, which enforces constrained traversal between a specific latent and the generated latent, guided by the real image distribution. To ensure consistency, the method spatially adjusts the noise to generate similar edits and employs the off-the-shelf SDXL model for denoising. For evaluation, the authors compared their method with Versatile Diffusion (VD), Paint-By-Example (PBE), IP-Adapter, and MimicBrush, focusing on image quality and semantic consistency. The proposed PIXELS achieved **91.71** and **5.412** in terms of CLIP-I Score (↑) and FID (↓)**,** respectively, outperforming VD (68.79, 18.174), PBE (78.67, 8.543), IP-Adapter (81.32, 7.958), and MimicBrush (84.95, 7.951). Similarly, in terms of semantic adherence and visual realism, PIXELS performed best, scoring **1.40** and **1.40,** respectively. Although the method lacks automation for segmentation map generation, it provides fine-grained, region-wise control over the editing process. Overall, PIXELS demonstrates an effective balance between quality, realism, and user control, marking a strong contribution to example-based image editing**[1]**.

Buburuzan et al. proposed a new method called MObI (Multimodal Object Inpainting). The system can add or replace objects in both camera images and lidar data in a realistic and controlled way. MObI uses a 3D bounding box instead of older 2D masks to place objects in the correct position and size within a scene. The author improved the Paint-by-Example (PbE) model by adding a diffusion model that works with both camera and lidar together. They also used cross-modal attention and a bounding box adapter to keep consistency between image and lidar data. The proposed system only needs one reference image of the object and can recreate it in the scene while keeping depth and lighting correct. The model was trained using the **nuScenes dataset** and uses CLIP to understand the object's features. The results showed that MObI makes objects look more realistic and better placed than older methods. It achieved **FID = 6.60**, **LPIPS = 0.115**, and **CLIP-I = 84.22**, which are better than the previous PbE model. Tests also showed that the added objects can be detected correctly by other detection models, proving the method's spatial accuracy. Although it does not perform as well in out-of-domain scenes, and may sometimes change the background. However, MObI still maintains realism, accuracy, and control much better than older 2D inpainting methods, making it a strong tool for creating synthetic driving data**[2]**.

Kim et al. proposed RAD (Region-Aware Diffusion Models) for fast and accurate image inpainting. The model RAD uses different noise schedules for each pixel, allowing it to focus only on the missing parts while keeping the rest of the image unchanged. The authors improved the standard diffusion model by adding spatial noise embedding, Perlin noise–based mask generation, and Low-Rank Adaptation (LoRA) for fine-tuning pretrained models efficiently. This setup makes the model both lightweight and effective. RAD performs inpainting by selectively adding noise only inside masked areas and then reconstructing them through a simple reverse process without any extra modules or complex loops.RAD was tested on FFHQ, LSUN Bedroom, and ImageNet datasets, where it achieved the best quantitative results compared to previous state-of-the-art methods. On the **FFHQ dataset**, RAD achieved an **FID of 22.1** and **LPIPS of 0.074**, outperforming MCG (FID 23.7, LPIPS 0.089) and RePaint (FID 25.7, LPIPS 0.093). On **ImageNet**, it reached **FID 47.0 and LPIPS 0.118**, while RePaint scored 54.0 / 0.177 and MCG 48.1 / 0.132. RAD also runs much faster, up to 100times quicker than RePaint and 15times faster than MCG, with an inference time of only 8.4 seconds on a 256×256 image. Although RAD needs some retraining and depends on the mask type, the authors reduced this problem using LoRA fine-tuning. Overall, RAD gives high image quality, fast results, and a simple design, making it an efficient and powerful model for image inpainting**[3]**.

Lugmayr et al. proposed RePaint, a novel inpainting technique that uses Denoising Diffusion Probabilistic Models (DDPM) to produce high-quality and diverse output images for any free inpainting form.RePaint leverages a pretrained unconditional DDPM and adapts it by conditioning the reverse diffusion process using the known regions of the image. The authors focused on this approach to avoid the need for mask-specific training, allowing the model to handle various types of masks during inference. The methods that have been implemented are **Pretrained Unconditional DDPM**, where the model is conditioned on the known image regions during the reverse diffusion process, **Resampling method** which is used to improve image harmonization and semantic coherence during the inpainting process. The model uses resampling within the reverse diffusion iterations and **No mask-specific training**, which is used to handle diverse and extreme masks. The model's flexibility is demonstrated through its ability to handle diverse and extreme mask types across datasets like **CelebA-HQ** and **ImageNet**. The two different masks have been used, such as thick and thin mask, which allows the authors to comprehensively assess RePaint's performance in a variety of inpainting challenges. RePaint outperforms all baseline methods **(AOT, DSI, ICT, LaMa)** on thin masks (e.g., "Super-Resolution 2x", "Alternating Lines"), with **LPIPS scores of 0.059 (CelebA-HQ)** and **0.134 (ImageNet)**. It achieves high texture fidelity and semantic alignment, with user votes ranging from **73.1% to 99.3%**. For thick masks (e.g., "Half Image", "Expand"), RePaint still performs well, though LaMa may outperform it in LPIPS for certain mask types, with **LPIPS** scores of **0.165 (CelebA-HQ)** and **0.304 (ImageNet).** Overall, RePaint delivers more realistic, semantically consistent inpainting with fewer artifacts, making it a leading method for free-form image inpainting **[4].**

Yang et al. proposed Uni-paint, a unified framework for multimodal inpainting based on pretrained Stable Diffusion that offers various modes of guidance, including unconditional, text-driven, stroke-driven, and exemplar-driven inpainting, as well as a combination of these modes. In order to overcome the limitations of single-modal guidance, the authors improved this method and formed a Uni-paint multimodal inpainting model. The authors use **Stable Diffusion** as the pretrained model, applying **masked finetuning** to adapt it for inpainting without requiring large-scale task-specific training datasets. They implement DDIM denoise sampling and introduce a masked attention control mechanism to prevent inpainted regions from spilling into the known areas. The model is evaluated quantitatively using metrics like **T2I** and **NIMA**. The Uni-paint framework outperforms other state-of-the-art methods in various inpainting tasks. For **text-driven** and **unconditional inpainting**, it achieves an **NIMA score** of **4.59** and a **T2I** score of 2**6.48**, slightly surpassing **SD-Inpaint (4.53 NIMA)**. In exemplar-driven inpainting, it excels with an **I2I score of 78.41** and **69.36% human preference** votes, outperforming **Paint-by-Example (21.27%)** and **TxtInv+BLD (9.36%)**. For stroke-driven inpainting, **Uni-paint** achieves **54.64% of human votes**, surpassing **SDEdit+BLD (45.36%**). The introduction of masked attention control ensures seamless inpainting with no boundary artifacts. Overall, Uni-paint provides superior flexibility, high-quality, and semantically meaningful results across all guidance modes, especially when compared to single-modal methods**[5]**.

Wasserman et al. proposed an outstanding image editing framework that reverses the inpainting process to add objects to images based on natural language instructions. They create a large dataset by removing objects from images using inpainting models and segmentation masks, then train a diffusion model to add objects back. The method uses a **Stable Diffusion (SD) model**, **Large Vision-Language Models (VLM),** and **Large Language Models (LLM)** to generate detailed instructions. The model outperforms existing methods, achieving state-of-the-art performance in object addition and general editing tasks, with improved **CLIP-I, DINO,** and **CMMD** scores. The Paint by Inpaint framework achieves exceptional performance, surpassing models like **IP2P** and **Hive** in object addition and image editing. It excels with **L1 (0.072)** and **L2 (0.025) metrics**, indicating high image consistency, and outperforms competitors in **CLIP-I (0.900)** and **DINO (0.852**) scores for semantic similarity. The **CMMD score (0.301**) further demonstrates its robustness. **Human evaluations** show **72.6%** preference for the model, highlighting superior edit faithfulness and quality. The PIPE dataset of 1 million image pairs enhances scalability and accuracy. So the authors overall demonstrate that training a diffusion model on the dataset leads to state-of-the-art performance in instruction-based image editing, proving the value of the PIPE dataset in achieving consistent and realistic image edits**[6].**

References

1. Das Biswas, S., Shreve, M., Li, X., Singhal, P., & Roy, K. (2025). PIXELS: Progressive image exemplar-based editing with latent surgery. *arXiv preprint arXiv:2501.09826*. https://arxiv.org/abs/2501.09826

2 . Buburuzan, A., Sharma, A., Redford, J., Dokania, P. K., & Mueller, R. (2025). *MObI: Multimodal object inpainting using diffusion models*. In CVPR 2025 Workshop on Data-Driven Autonomous Driving Simulation (DDADS). Computer Vision Foundation.

https://openaccess.thecvf.com/content/CVPR2025W/DDADS/papers/Buburuzan_MObI_Multimodal_Object_Inpainting_Using_Diffusion_Models_CVPRW_2025_paper.pdf

3. Kim, S., Suh, S., & Lee, M. (2025). *RAD: Region-aware diffusion models for image inpainting*. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Computer Vision Foundation. https://openaccess.thecvf.com/content/CVPR2025/papers/Kim_RAD_Region-Aware_Diffusion_Models_for_Image_Inpainting_CVPR_2025_paper.pdf

4. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., & Van Gool, L. (2022). RePaint: Inpainting using denoising diffusion probabilistic models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, 10442-10452. https://openaccess.thecvf.com/content/CVPR2022/html/Lugmayr_RePaint_Inpainting_Using_Denoising_Diffusion_Probabilistic_Models_CVPR_2022_paper.html

5. Yang, S., Chen, X., & Liao, J. (2023, October). Uni-paint: A unified framework for multimodal image inpainting with pretrained diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia* (pp. 3190–3199). https://doi.org/10.1145/3581783.3612200

6. Wasserman, N., Rotstein, N., Ganz, R., & Kimmel, R. (2025). Paint by inpaint: Learning to add image objects by removing them first. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18313–18322. https://openaccess.thecvf.com/content/CVPR2025/html/Wasserman_Paint_by_Inpaint_Learning_to_Add_Image_Objects_by_Removing_CVPR_2025_paper.html