Scene Embeddings via Contrastive Learning

Ryan Bergamini
Whiting School of Engineering
Johns Hopkins University
Baltimore, MD
rbergam1@jh.edu

Dr. Erhan Guven
Whiting School of Engineering
Johns Hopkins University
Baltimore, MD
eguven2@jhu.edu

Vy Vu
Whiting School of Engineering
Johns Hopkins University
Baltimore, MD
vvu5@jhu.edu

***Abstract*—This research explores scene-aware inpainting through a novel contrastive learning approach, focusing on learning scene embeddings to guide context-consistent image generation. This research investigates how different neural architectures—convolutional, transformer-based, and hybrid models—perform in generating embeddings that capture scene-specific information. Using images of chairs from the COCO dataset, this research contributes a novel benchmark by masking chairs and generating hard negative examples with a state-of-the-art inpainting model. This research proposes a modified Bootstrap Your Own Latent (BYOL) architecture for unsupervised learning of scene representations. The experiments reveal limitations in the effectiveness of current contrastive learning paradigms for this task, with minimal differentiation between positive and negative pair similarities. These findings highlight challenges in learning meaningful scene representations for context-aware image generation.**

***Index Terms*—Contrastive Learning, Image Inpainting, Scene Embeddings, Generative Adversarial Networks, Computer Vision, Deep Learning, Transformer Networks, Convolutional Neural Networks, Image Generation, Unsupervised Learning**

I. Introduction

The purpose of this research is to make progress towards the ultimate of objective of realistically inserting images of existing objects into scenes. The field of image editing and manipulation via Artificial Intelligence has already delivered numerous advances in generative models (e.g. DALL-E [1], Stable Diffusion [2]). These generative models prove useful for artistic expression or general representation of a semantic idea. The challenge is to adjust these models to generate existing images in context. This research reviews the history of generative artificial intelligence and seeks to leverage recent progress in contrastive learning to generate images more consistent with existing scenes.

II. Related Work

Below is a summary of existing efforts for Image Editing, Manipulation, and Interpretation with Artificial Intelligence.

A. Object-Based Image Editing

Before a specific object within an image can be edited, it must first be identified and segmented. The foundation for in-context image editing lies in early object-based segmentation techniques. Barrett & Cheney [3] introduced one such method using a watershed algorithm to construct a triangular mesh over segmented objects, enabling localized manipulation. The watershed algorithm treats pixel intensities as elevations on a topographic map and simulates the flooding of basins from local minima, forming boundaries where different ”watersheds” converge.

Recent advances in object segmentation have dramatically improved the flexibility and generalization of object-aware editing systems. Notably, Meta AI’s *Segment Anything Model (SAM)* (Kirillov et al., 2023) uses a Vision Transformer (ViT) backbone and is trained on the SA-1B dataset, which includes over 11 million images and 1.1 billion segmentation masks. SAM enables zero-shot and one-shot segmentation through prompt-able inputs such as points, bounding boxes, and masks, making it a highly versatile tool for downstream applications in semantic image editing [4].

B. Generative Adversarial Networks

Once an object is detected, the next challenge is generating replacement content that is both semantically appropriate and stylistically consistent with the surrounding image. Semantic control determines *what* type of object is generated (e.g., a tree, a bicycle), while stylistic control governs *how* that object appears in the context of the scene (e.g., lighting, texture, color). The introduction of Generative Adversarial Networks (GANs) by Goodfellow et al. [5] marked a turning point in generative modeling by enabling high-quality image synthesis that was both semantically and stylistically realistic. A GAN consists of two neural networks—a generator and a discriminator—engaged in a two-player minimax game. The generator maps random noise vectors or latent encodings to synthetic images, while the discriminator learns to distinguish real images from those generated. Through adversarial training, the generator incrementally improves its output to better mimic the true data distribution, resulting in increasingly realistic imagery.

The core architecture of both the generator and discriminator can vary based on the task. Most early implementations employed convolutional neural networks (CNNs), particularly in image domains, due to their spatial inductive biases and translation invariance. For instance, the Deep Convolutional GAN (DCGAN) introduced by Radford et al. (2016) used a fully convolutional architecture with batch normalization and leaky ReLU activations, setting a precedent for many subsequent image-based GAN models [6]. Later variants experimented with attention mechanisms and transformer architectures to capture long-range dependencies, especially in high-resolution generation tasks.

Several architectural extensions have been proposed to enhance semantic control over generated content. A key milestone in this direction was the introduction of Conditional GANs (cGANs) by Mirza and Osindero [7], which modified the original GAN framework to incorporate auxiliary information such as class labels or attributes. In a cGAN, both the generator and discriminator receive conditioning variables—typically concatenated with the noise vector (in the generator) or the image (in the discriminator)—allowing the model to learn mappings from specific inputs to controlled outputs. This enabled a fundamental shift from purely random image generation to targeted synthesis, where users could guide the generative process.

Building on the advances of cGAN, CycleGAN and StyleGAN enabled greater semantic control from sources of existing information. **CycleGAN** [8] introduced a framework for unpaired image-to-image translation by incorporating a cycle-consistency loss. CycleGAN consists of two generators—one that translates an image from Domain A to Domain B and the other vice versa. Cycle-consistency loss ensures that translating from A to B and back again to A yields a result similar to the original, enabling the model to learn meaningful mappings between domains without requiring aligned image pairs.

StyleGAN [9], in contrast, offered fine-grained control over the style and attributes of generated images by introducing a mapping network and adaptive instance normalization (AdaIN) layers, allowing the disentanglement of coarse and fine features across latent space. StyleGAN models take advantage of the insight that coarser details such as pose and orientation are stored in early layers, while finer details such as freckles or pores are controlled by higher-resolution layers. A StyleGAN model learns to interpret an intermediate latent vector $\mathbf{w}$, derived from a sampled noise vector $\mathbf{z}$, to achieve semantic control over each layer. Gaussian noise is added independently at each resolution to introduce stochastic detail. However, since these fine-grained details are randomly injected, users must often sample multiple generations to obtain a desired effect such as freckles or texture.

C. Incorporating Attention into GANs

To improve the spatial awareness and contextual fidelity (i.e. the stylistic attributes) of generated images, researchers have incorporated various forms of attention mechanisms into GAN architectures. **SPA-GAN** [10] introduces a spatial attention mechanism designed to preserve critical regions of the image during translation. It operates by learning attention maps that weight spatial features, helping the generator focus on structurally important areas (e.g., edges, facial landmarks). This attention is applied asymmetrically—once per domain—to emphasize spatial consistency in domain translation.

In contrast, **Dual Attention GANs** [11] adopt a more complex mechanism that integrates both *spatial* and *channel-wise attention* to generate photorealistic images from existing segmented images. Their model first computes channel-wise feature statistics, aggregates information across channels, and then re-applies attention weights to each feature map. This dual attention pipeline allows the model to reason not only about ”where” to focus but also ”what” semantic information to enhance, enabling fine-grained feature refinement across the entire image. While SPA-GAN emphasizes local preservation of image structure during translation, Dual Attention GANs aim to optimize both structural coherence and semantic salience at multiple scales.

D. High-Resolution and Context-Aware Editing

Pix2PixHD [12] introduced a framework for high-resolution image synthesis and semantic manipulation using conditional GANs. Their architecture incorporates multiple discriminators operating at different spatial resolutions, allowing the model to enforce both global coherence and local realism in the generated images. A key innovation in their system is the use of a context-aware feature encoding strategy, which enables the model to distinguish between different instances of semantic categories (e.g., distinguishing two cars or buildings within the same label class).

To support interactive editing, the authors developed an interface that allows users to manipulate semantic label maps and instance-level encodings. To enable object-specific control, the model uses an *autoencoder* to extract feature vectors for individual object instances. These vectors are then clustered using *K-means*, producing a set of interpretable semantic feature clusters. During editing, users can assign or swap cluster labels to alter the appearance of specific objects within the scene, providing fine-grained control over textures, colors, and structural variation. This clustering-based interface enables semantically meaningful edits without requiring manual pixel-level supervision.

RealFill [13], a recent advance in inpainting, addresses the challenge of generating plausible image content within masked regions using both local and global contextual cues. Unlike traditional GAN-based inpainting approaches, RealFill is built on a *fine-tuned diffusion model*, which generates images through a denoising process that gradually transforms random noise into structured content. Diffusion models differ fundamentally from GANs: rather than relying on a discriminator for adversarial feedback, they use a probabilistic forward-and-reverse process to iteratively refine the image. This iterative nature enables them to capture both stylistic detail and global semantic coherence more effectively than single-pass GAN architectures.

RealFill further enhances this process by incorporating a few reference images of the same scene. These references provide high-level cues about lighting, object geometry, and semantic layout, allowing the model to generate content that is consistent with the broader scene. During training, the model is fine-tuned on just a few views of a specific environment, enabling it to internalize the visual structure and conditions of that scene.

GANPaint [14] introduces an interactive image editing system that allows users to add or remove semantic objects (e.g., windows, trees) by directly modifying the internal activations of a pretrained GAN. A core innovation of GANPaint is its use of *editing masks*, which spatially constrain where in the image a modification should occur. When a user selects a region to edit—such as brushing over a wall to add a window—this interaction produces a binary mask that identifies the specific spatial locations within a target feature map where the change should be applied.

GANPaint first identifies a set of *causal units*—individual channels in an intermediate layer of the GAN generator—that are strongly associated with the presence of specific objects. This association is determined through correlation analysis and validated using ablation and insertion tests. During editing, the binary mask is used to selectively modify the activations of these causal units only at the masked spatial locations. For example, to insert a window, GANPaint activates the relevant feature channels, but only within the masked region, leaving the rest of the generator's internal representation unchanged. This approach ensures that the modification is localized and semantically meaningful, while preserving the appearance and structure of the unedited parts of the image. By operating directly on the generator's internal representations, GANPaint achieves precise object-level control without needing to modify the input latent vector or retrain the model.

E. Visual Object Embeddings via Contrastive Learning

A related field to image manipulation is image encoding. In the SimCLR [15], researchers developed a methodology to encode images in a self-supervised manner (i.e. with no human labeling intervention) via contrastive learning. Contrastive learning generates embeddings for images via training on positive pairs - images that should be similar - and negative pairs - images that should be different. Contrastive learning generates the embeddings that are then used for downstream classification tasks.

The SimCLR method trains two encoding networks on both positive and negative pairs while the Bootstrap Your Own Latent (BYOL) [16] approach takes a different approach that eliminates the need for explicit negative pairs. BYOL consists of three key components:

1. **Encoder**: The backbone network (such as ResNet or Vision Transformer) that extracts high-dimensional features from input images. The encoder maps raw images to representation vectors.

2. **Projector**: A multi-layer perceptron (MLP) that transforms the encoder's output into a latent space where the contrastive learning occurs. This component helps to create a representation space better suited for the contrastive task.

3. **Predictor**: An additional MLP applied only in the online network branch that tries to predict the target projection vector. The predictor adds an asymmetry between the two branches that prevents collapse.

BYOL uses two networks: an online network (containing all three components) and a target network (containing only the encoder and projector). The target network's parameters are an exponential moving average (EMA) of the online network's parameters, which provides training stability. The lack of negative pairs in BYOL makes it more computationally efficient and allows it to avoid potential pitfalls when selecting appropriate negative examples.

F. Novel Scene-Aware Inpainting Approach

While existing research has made significant progress in both contrastive learning and context-aware image generation, there remains a gap in leveraging scene embeddings to guide the inpainting process. This research bridges this gap by introducing a novel approach that combines contrastive learning-based scene representations with generative adversarial networks. Previous works like GANPaint [14] control internal activations to preserve scene coherence but are limited to predefined semantic concepts and require direct manipulation of the generator's feature maps. In contrast, the proposed EmbeddingGAN model generates complete image outputs while applying the original masked image as a guarantee that unedited regions remain unchanged. This approach offers greater flexibility while maintaining high fidelity to the original scene.

This work leverages and extends several key concepts from recent advances in image generation:

1. The EmbeddingGAN model incorporates the attention mechanisms for incorporating scene-embedding information, which have proven effective at maintaining spatial and semantic coherence as demonstrated in SPA-GAN [10] and Dual Attention GAN [11]

2. The EmbeddingGAN model replaces semantic conditioning from conditional GANs [7] with learned scene embeddings derived through contrastive learning.

3. The proposed modified BYOL approach extends BYOL's self-supervised representation learning approach [16] to specifically focus on scene context rather than object recognition, creating embeddings that capture the contextual relationships within an image.

4. Unlike GANPaint [14], which selectively controls which generator activations are turned on or off to maintain scene consistency, the EmbeddingGAN takes a more direct approach by generating a complete new image and then pasting the original masked portions over the generated content. This ensures perfect preservation of unedited areas while allowing the model to focus exclusively on generating contextually appropriate content for the masked regions.

5. Unlike RealFill [13], which requires multiple reference images of the same scene, this research's approach aims to infer context from a single image, making it more practical for real-world applications.

III. Methodology

The research applies contrastive learning to scenes and masks to learn scene specific embeddings that would be used in a Generative Adversarial Network to create more realistic scene fillings. The research attempts to determine if a transformer-based, convolutional neural network-based, or hybrid encoder architecture is more effective at generating scene embeddings.

A. Data Preparation

For this research, only COCO images from train2017 that contain chairs (approximately 12.2K images) were used due to the ultimate intended application of this research to quickly visualize what a specific chair would look like in an existing scene. Masks of the chairs were generate with COCO annotations. To provide the contrastive learning model with hard examples that support learning, the LaMa [19] model was then used to fill the mask gap in the original image. The dataset of 12.2K images - including the original image, the masked images, the image masks, the precomputed masks of chairs, the infilled image and the precomputed masks of the infilled portion of the image is posted publicly on Kaggle https://www.kaggle.com/datasets/ryanbergamini/inpainted-chairs-from-coco-2017/.

B. Contrastive Learning

For the research, a modified Bootstrap Your Own Latent contrastive learning architecture is used. The standard BYOL approach does not make use of negative pairs due to the computational effort to generate negative training pairs for each patch. Since the Stable Diffusion inpainted masks are pre-generated for the dataset, this work adapts the BYOL-loss function by using the similarity between two negative pairs (a number that is closer to 1 when similarity is high) - as opposed to the reciprocal of the similarity for positive pairs (a number that is 0 when similarity is high).

The modified BYOL approach still consists of the three key components: an encoder, a projector, and a predictor, with both online and target networks. For data augmentation, the modified BYOL only uses random crop selection. The traditional methods that modify the color distributions to encourage more semantic learning are not used since the purpose of this research is to embed the scene and not the semantics of an image.

1) Network Architectures: This work implement and evaluate three types of encoders:

- **CNN-based encoder:** Built on a ResNet50 [17] backbone pre-trained on ImageNet. This encoder extracts convolutional features from input images, capturing spatial hierarchies and local patterns.
- **Transformer-based encoder:** Utilizes the Vision Transformer (ViT-Base) architecture [18] with patch size 16×16, pre-trained on ImageNet. This encoder processes images as sequences of patches and leverages self-attention mechanisms to capture long-range dependencies.
- **Hybrid encoder:** Combines both CNN and transformer features by extracting and concatenating features from both architectures, leveraging the strengths of each approach.

In all three cases, the encoder outputs are fed into a projector network, which is a Multi-Layer Perceptron (MLP) with three linear layers, batch normalization, and ReLU activations between the layers. The projector maps the encoder's high-dimensional features to a 256-dimensional embedding space. Finally, a predictor MLP with a similar structure to the projector is applied to the online branch's projection, attempting to predict the target network's projection.

The quality of the contrastive learning embedding space is assessed by the mean similarity between images and the cropped object, the image and the inpainted object, and the inpainted object and the cropped object.

2) Training Details: All BYOL models were trained with the following hyperparameters:

- Batch size: 64
- Learning rate: 0.0003
- Optimizer: Adam with weight decay 1e-6
- EMA decay rate for target network: 0.99
- Training epochs: 10

A batch size of 64 had to be used due to hardware memory limitations. All models were also only trained on 4000 randomly sampled images from the dataset due to the same limitations.

C. Embedding GAN Architecture

The Generative Adversarial Network would use the trained modified BYOL encoder and projector to create the embedding for the masked scene. The embedding of the masked scene is then provided to a cross-attention module with a latent vector. The latent vector would then be used to calculate the embedding loss from applying the same modified BYOL encoder and projector. The embedding loss will be balanced with the standard GAN loss and reconstruction loss to encourage the generator to create realistic image that is also true to the scene embedding.

The full code repository which includes model implementations is available at https://github.com/rberg27/scene-embedding-gan.

D. Metrics

Which type of encoder - transformer, CNN, or hybrid - is more effective will be determined by the similarity metrics between positive and negative pairs as well as a qualitative comparison between the images generated from the Embedding GAN that makes use of the modified BYOL architecture. The images produced by the EmbeddingGAN will also be compared to the results of a Vanilla GAN with the masked image pasted on top of the Vanilla GAN output (hereby refered to as a StencilGAN). The images inpainted with the LaMa model is also included for reference of the Embedding GAN performance against an existing solution.

IV. Research Hypothesis

The transformer-based modified BYOL encoder would generate scene embeddings with greater positive pair similarity and negative pair dissimilarity as well as generate more realistic images. SPA-GAN demonstrated that the attention model is more effective at generating context aware images than a convolutional network - therefore it is likely the same will apply for generating scene embeddings via contrastive learning.

V. Experimental Results and Analysis

A. Scene Embedding Similarities

Table I documents the mean cosine similarity between masked images and the cropped objects, the masked images and inpainted objects, and the masked images and filled objects

TABLE I: Cosine Similarity of Scene Embeddings

BYOL Encoder Type	Pos-Pair Similarity	Neg-Pair Similarity
CNN	0.981	0.957
Transformer	0.959	0.944
Hybrid	0.978	0.950

Each of the BYOL models with different encoder architectures were trained for 10 epochs on 4K of the 12.2K images with a batch size of 64. Although the positive pairs have higher similarity for each encoder type, the difference in the average negative pair similarity is not significant. All three encoder methods produced a higher similarity between positive pairs than for negative pairs. Although the difference is not significant, the fact that all three produced greater similarity suggests an architecture could be designed to make that difference significant.

B. EmbeddingGAN Output

Figure 1 and Figure 2 attached to this research present output for the EmbeddingGAN compared to Stable Diffusions-based LaMa model inpainting and the StencilGAN as well as the comparison of EmbeddingGAN quality for each of the three types of encoders. No model produces realistic results. Factors that could have contributed to the lackluster results include that no encoder produced significant difference between positive and negative pairs in the scene embedding space. The lack of difference suggests no meaningful scene specific information is encoded in the scene embeddings used to inform the EmbeddingGAN generation. This is evident based on the similar output between the StencilGAN and EmbeddingGAN - where the StencilGAN was provided no scene embedding.

The lack of realistic results from the LaMa model suggests that the cause for the lack of results is due to the complexity of the task. The Stable Diffusion has proven to be an effective image inpainting model - the output of the Stable Diffusions based LaMa model were marginally more coherent than the StencilGAN and EmbeddingGAN output. An area of future research is including a semantic embedding in a revised EmbeddingGAN architecture similar to that of cGAN [7] to inform improved infilling.

VI. Conclusion

Using contrastive learning to create scene specific embeddings has thus far been unsuccessful. No encoder method for creating scene embeddings were distinguishable as the superior method. The lack of differentiation between mean positive-pair similarity and negative-pair similarity did not provide any observable advantage for the EmbeddingGAN model over the baseline of a VanillaGAN output that was provided no scene specific information.

Despite the insignificant results from this research, the novel concept of creating scene-specific embeddings to complete masked images deserves to be further explored. Further research should explore how to build a contrastive learning model can create meaningful difference between positive and negative pairs in a dataset. Once such a model that can encode meaningful scene-specific embeddings is created, the EmbeddingGAN has the potential to help realistically insert an existing image into a scene.

Acknowledgment

References

[1] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I. (2021). Zero-Shot Text-to-Image Generation. arXiv preprint arXiv:2102.12092.

[2] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B. (2021). High-Resolution Image Synthesis with Latent Diffusion Models. arXiv preprint arXiv:2112.10752.

[3] Barrett, W. A., Cheney, A. S. (2002). Object-based image editing. ACM Transactions on Graphics (TOG), 21(3), 777-784.

[4] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, L., Gustafson, T., Xiao, T., Whitehead, S., Cardenas, A., Berg, T. L., et al. (2023). Segment Anything. arXiv preprint arXiv:2304.02643.

[5] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. (2014). Generative Adversarial Networks. Advances in Neural Information Processing Systems, 27, 2672-2680.

[6] Radford, A., Metz, L., Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

[7] Mirza, M., Osindero, S. (2014). Conditional Generative Adversarial Nets. arXiv preprint arXiv:1411.1784.

[8] Zhu, J.-Y., Park, T., Isola, P., Efros, A. A. (2017). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2223-2232.

[9] Karras, T., Laine, S., Aila, T. (2019). A Style-Based Generator Architecture for Generative Adversarial Networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 4401-4410.

[10] Emami, H., Aliabadi, M. M., Dong, W., Chinmay, R. (2019). SPA-GAN: Spatial Attention GAN for Image-to-Image Translation. IEEE Transactions on Multimedia.

[11] Tang, H., Qi, H., Xu, D., Wang, P., Sebe, N. (2020). Dual Attention GANs for Semantic Image Synthesis. Proceedings of the 28th ACM International Conference on Multimedia, 1994-2002.

[12] Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., Catanzaro, B. (2018). High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 8798-8807.

[13] Kulal, S., Yin, H., Stark, A., Hu, S. X., Tenenbaum, J. B., Isola, P., Chan, E., Klein, R. (2023). RealFill: Reference-Driven Generation for Authentic Image Completion. arXiv preprint arXiv:2309.16668.

[14] Bau, D., Zhu, J.-Y., Strobelt, H., Zhou, B., Tenenbaum, J. B., Freeman, W. T., Torralba, A. (2019). GAN Paint: Semantic Image Manipulation with a Generative Image Prior. ACM Transactions on Graphics, 38(4), 1-11.

[15] Chen, T., Kornblith, S., Norouzi, M., Hinton, G. (2020). A Simple Framework for Contrastive Learning of Visual Representations. Proceedings of the 37th International Conference on Machine Learning, 1597-1607.

[16] Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M. (2020). Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning. Advances in Neural Information Processing Systems, 33, 21271-21284.

[17] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770-778.

[18] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations (ICLR).

[19] Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V. (2022). LaMa: Resolution-robust Large Mask Inpainting with Fourier Convolutions. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2149-2159.

[20] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10684-10695.

[21] Elyan, E., Jamieson, L., Ali-Gombe, A. (2020). Deep learning for symbols detection and classification in engineering drawings. Neural Networks: The Official Journal of the International Neural Network Society, 129, 91-102. https://doi.org/10.1016/j.neunet.2020.05.025

[22] Lin, Y., Ting, Y., Huang, Y., Cheng, K., Jong, W. (2023). Integration of Deep Learning for Automatic Recognition of 2D Engineering Drawings. Machines, 11(8), 802. https://doi.org/10.3390/machines11080802

[23] Moreno-García, C., Elyan, E., Jayne, C. (2018). New trends on digitisation of complex engineering drawings. Neural Computing and Applications, 31, 1695-1712. https://doi.org/10.1007/s00521-018-3583-1

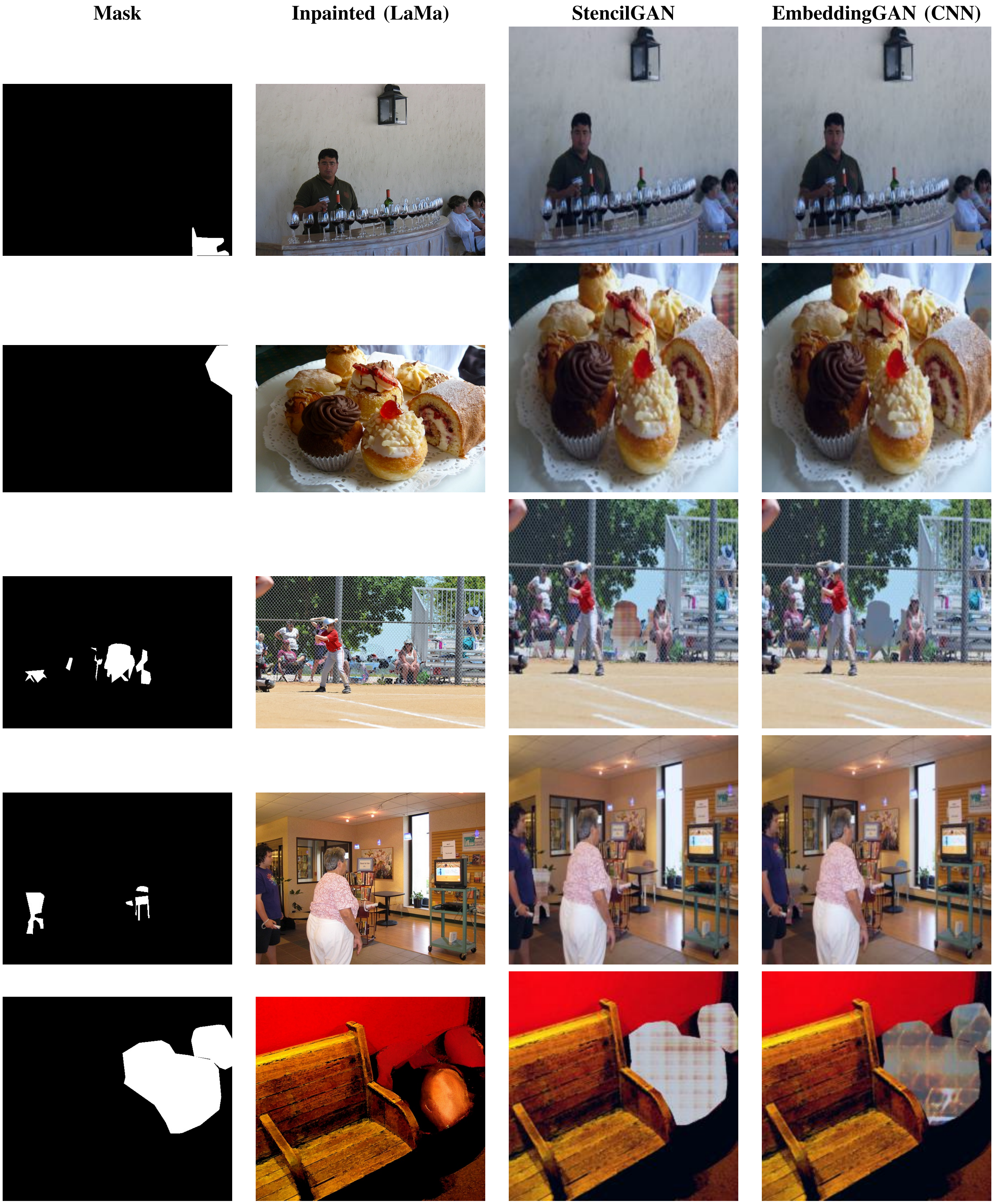

Fig. 1: Comparison of inpainting methods from Standard Diffusion, StencilGAN (a Vanilla GAN baseline), and the Embedding GAN (with a CNN encoder). No model produced strong results.

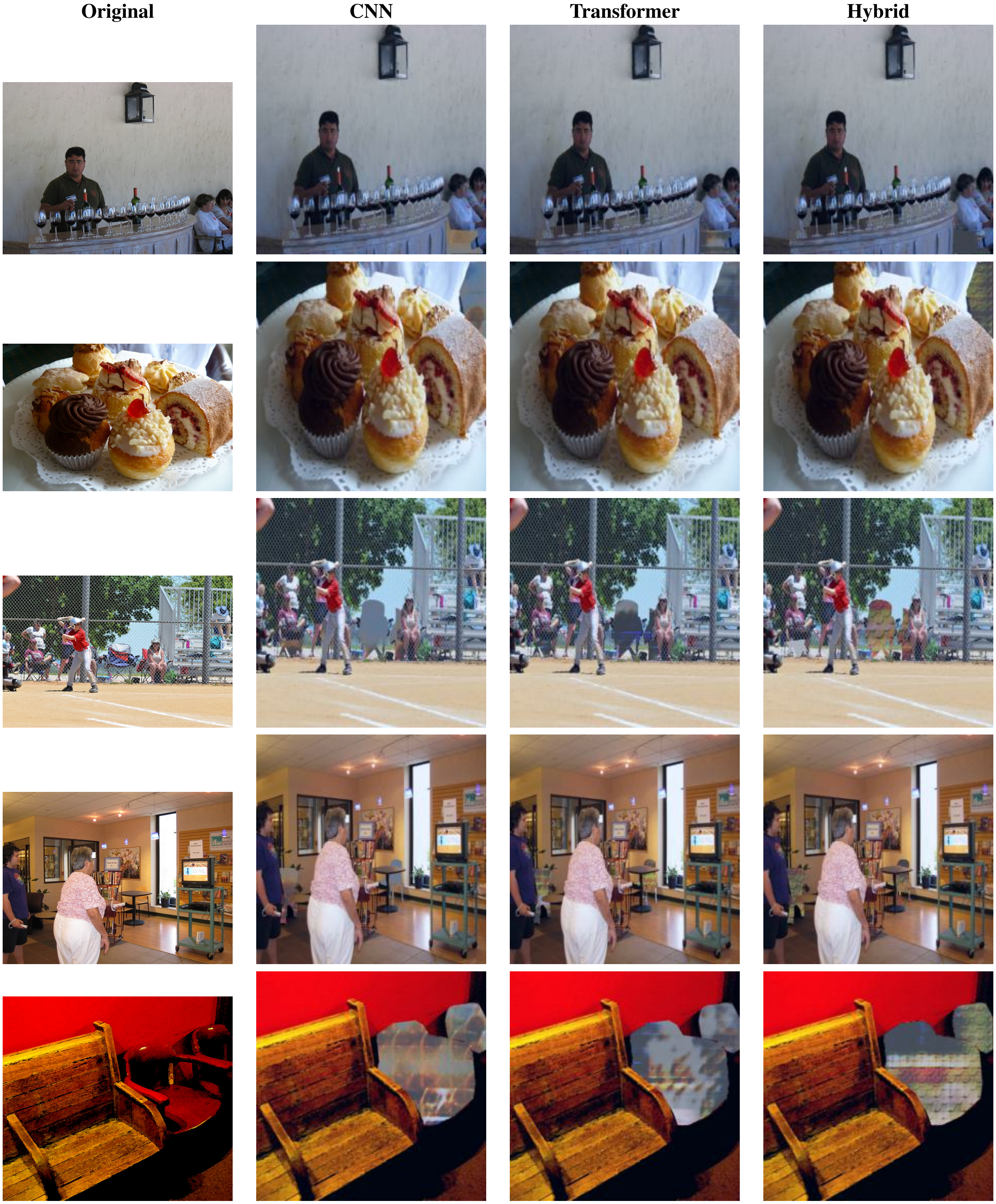

Fig. 2: Comparison of inpainting results using different embedding approaches. From left to right: original image (with chair present), CNN-based encoder generated result, transformer-based encoder generated result, and hybrid encoder approach generated result. Each row represents a different test case from the COCO dataset.