Self-Supervised Reconstruction of Shuffled Images Using Encoder–Decoder Networks

Kavana Manvi Krishnamurthy
kmanvikr@depauledu

Shardul Janaskar
sjanaska@depaul.edu

Shams Nahid Kotnur
skotnur@depaul.edu

Abstract*— Reconstructing images from shuffled patches is a challenging task that simulates real-world problems such as artifact restoration, corrupted document recovery, and forensic analysis. In this study, we propose a self-supervised learning approach for solving visual jigsaw puzzles, using a convolutional encoder–decoder architecture. The network learns to reassemble randomly shuffled image patches without supervision by minimizing a combination of pixel-level reconstruction loss and a spatial consistency term. We conduct experiments on the Places365 dataset using a controlled patch-shuffling protocol focusing exclusively on easy level where in two patches were shuffled. Our results demonstrate strong reconstruction performance at the easy level, as quantified by high Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR) values. Training curves show consistent improvement across evaluation metrics, validating the effectiveness of our method. This research serves as a foundation for exploring more advanced spatial reasoning in vision systems, with future extensions targeting superpixel-based patching and more complex puzzle configurations.

Keywords*— Image Reconstruction, Jigsaw Puzzle, Deep Learning, Self-Supervised Learning, CNN, Patch Reordering, Spatial Consistency

INTRODUCTION

The reconstruction of images from shuffled fragments presents an intricate challenge in the domain of computer vision, demanding more than conventional pixel-wise restoration. This task simulates practical scenarios such as the restoration of degraded photographs, the reassembly of fragmented documents, and artifact reconstruction in archaeology. Successfully addressing this problem requires models not only to restore individual pixel values but also to capture the underlying spatial dependencies and contextual relationships among image components.

In this study, we approach the image puzzle problem using a self-supervised learning strategy. Rather than relying on labeled data, the model learns to reconstruct images by recovering their original structure from versions where some patches have been randomly rearranged. This encourages the network to develop an understanding of spatial composition without explicit supervision.

Each image is first resized to 256×256 pixels and divided into a 4×4 grid, resulting in 16 equally sized patches. Based on the defined difficulty level easy, medium, or hard, 2, 4, or 6 patches are randomly shuffled, ensuring none remain in their original positions. The shuffled image serves as input, while the original image acts as the reconstruction target. Figure 1 illustrates sample inputs with shuffled patches and the corresponding outputs generated by our autoencoder model after training for 50 epochs.

.

Fig. 1. Input is a reshuffled image and the output is the reassembled image.

LITERATURE REVIEW

[1] Freeman, W.T., Gardner, W.A. "Shape-based puzzle solving for archaeology," Computer Vision, 1995. Freeman and Gardner's foundational work pioneered the computational reassembly of fragmented objects, with a focus on archaeological artifacts. Their approach emphasized shape-based matching, using geometric cues such as contour continuity and edge alignment to pair fragment boundaries. While highly effective for structured domains with well-preserved edges, this heuristic-driven strategy is less suitable for puzzles where physical boundaries are eroded or indistinct. Nonetheless, their contributions laid essential groundwork, introducing fundamental principles of geometric compatibility that continue to influence puzzle-solving research.

[2] Cho, T.S., Avidan, S., Freeman, W.T., "A Probabilistic Image Jigsaw Puzzle Solver," IEEE TPAMI, 2010. Cho et al. Cho and colleagues introduced a significant shift in jigsaw puzzle solving by transitioning from shape-based strategies to image-based reasoning. Their probabilistic model framed puzzle reconstruction as a maximum likelihood problem, estimating patch adjacency based on color consistency and gradient flow across edges. This shift allowed the method to scale effectively to larger puzzles and handle a wider variety of natural images. However, its reliance on low-level visual features limited its effectiveness in semantically complex scenes, setting the stage for deeper, learning-based approaches to capture more abstract spatial cues.

[3] Noroozi, M., Favaro, P., "Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles," CVPR, 2016. A landmark in self-supervised learning, Noroozi and Favaro reframed jigsaw solving as a classification task embedded in a deep learning framework. By randomly shuffling image patches and training a convolutional neural network to predict the correct arrangement, they demonstrated that spatial configuration prediction could drive the learning of transferable visual representations. Crucially, this approach required no labeled data and yielded features beneficial for downstream tasks like object detection and scene classification, helping to bridge the gap between supervised and unsupervised visual learning.

[4] JigsawGAN: Auxiliary Learning for Solving Jigsaw Puzzles with Generative Adversarial Networks (Lee et al., 2020) introduces *JigsawGAN*, a novel framework that leverages GANs to solve jigsaw puzzles by jointly learning two tasks. JigsawGAN introduces a hybrid framework that leverages both generative and discriminative learning to enhance puzzle solving. The model simultaneously learns to reorder shuffled patches and reconstruct coherent images, with the generator tasked with image synthesis and the discriminator validating permutation plausibility. This auxiliary generation objective enables the model to develop richer spatial and semantic understanding than classification alone. Notably, the GAN-based structure shows robust performance in high-complexity settings with many permutations, highlighting the synergy between generative modeling and self-supervised learning for spatial tasks.

[5] Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A., "Context Encoders: Feature Learning by Inpainting," CVPR, 2016. Pathak and collaborators proposed context encoders, which extend the idea of spatial reasoning into generative territory through image inpainting. By requiring a model to fill in missing regions based on surrounding context, the method implicitly teaches the network to capture global image structure and semantics. While differing from jigsaw tasks in formulation, this approach shares the underlying philosophy of using spatial gaps as supervision. The resulting representations proved highly transferable, reinforcing the effectiveness of generative pretext tasks in learning visual features.

[6] Solving Jigsaw Puzzles with Eroded Boundaries (Bridger et al., CVPR 2020) tackles a harder variant of the jigsaw puzzle problem where pieces have eroded boundaries. Bridger and colleagues addressed a challenging variant of the jigsaw problem involving eroded patch boundaries, where traditional edge-based alignment fails. Their approach employs a GAN-based inpainting model to fill the gaps between patches, using the quality of reconstructed transitions as a proxy for adjacency. The discriminator's feedback serves both as a loss and as a classifier for neighbor prediction. This method proves effective even in large puzzles (70–150 pieces), outperforming traditional techniques by leveraging learned structural cues. It offers valuable insights for scenarios involving occlusion, degradation, or missing data.

[7] Yu, A., Römer, K., "Solving Jigsaw Puzzles with Linear Programming," ECCV Workshops, 2016. Yu and Römer. In contrast to data-driven and heuristic approaches, Yu and Römer framed jigsaw solving as a formal optimization problem using linear programming. Their model maximized patch compatibility based on Mahalanobis gradient continuity, with global constraints ensuring consistent tile placement. They further introduced a two-phase pipelineinitial layout estimation followed by local refinement to improve accuracy. Although computationally intensive for large puzzles, this framework demonstrated that mathematical optimization can achieve high reconstruction fidelity when patch relationships are precisely quantified.

[8] Paumard, T., Kania, K.D., Ba, S.O., Pinquier, J., & Rouas, J.L., "Solving Jigsaw Puzzles with Combined Deep and Graphical Models," arXiv:2008.12959, 2020. Paumard and collaborators presented a hybrid model combining deep learning with graphical inference to improve robustness under noisy or occluded conditions. Their system extracts feature-rich patch embeddings using a ResNet-18 backbone, then builds a compatibility graph over patch pairs. An iterative message-passing algorithm enforces global coherence, enabling more accurate arrangements than greedy or classification-based models. This integration of CNNs and probabilistic graphical models offers a compelling strategy for tackling complex spatial reasoning tasks with structured uncertainty.

[9] Saharia, C., Ho, J., Salimans, T., Chan, W., Fleet, D.J., Norouzi, M. "ImageBART: Bidirectional Context with Multinomial Diffusion for Autoregressive Image Generation," arXiv, 2022. Saharia et al. propose ImageBART merges the concepts of bidirectional context modeling and diffusion-based autoregression for structured image generation. Inspired by language models like BART, it enables spatial understanding in both forward and backward directions, a distinct advantage for jigsaw-style tasks involving spatial reordering. The multinomial diffusion process helps recover high-fidelity details even from heavily corrupted inputs. While designed for generative applications, the architectural innovations and training objectives hold strong relevance for self-supervised tasks that rely on reconstructing or rearranging visual content.

METHODOLOGY

This work addresses the challenge of reconstructing a spatially consistent image from a shuffled version, formulated as a self-supervised learning task. The system is trained without explicit labels by using the original unshuffled image as the reconstruction target. The core stages in our methodology include data preparation, network design, loss formulation, and training strategy.

A. Dataset

The experiments in this study are done using the Places365 dataset, a large-scale scene-centric image dataset designed for benchmarking models on visual understanding of natural and man-made environments. It contains more than 1.8 million training images spanning 365 distinct scene categories, including both indoor and outdoor settings like forests, offices, markets, and residential areas.

For this project, we utilize a subset of 10,000 images sampled from the official validation split. All images are resized to a uniform resolution of 256×256 pixels to ensure consistency in patch extraction and model input dimensions. This resizing also facilitates efficient memory utilization and consistent patch decomposition.

To simulate the puzzle-solving task, each image is partitioned into a 4×4 grid of 64×64 patches, resulting in 16 non-overlapping segments. Depending on the difficulty level, a fixed number of these patches are randomly shuffled, while the original image is retained as the reconstruction target. For the current implementation, only the easy difficulty setting is considered, in which two patches are permuted per image.

The diversity of scenes in Places365 provides varied contextual information and visual complexity, making it a suitable benchmark for learning spatial reasoning and structural alignment in image reconstruction tasks.

B. Preprocessing Pipeline

All input images are uniformly resized to 256×256 pixels to maintain a fixed input resolution, enabling consistent division into square patches. Each image is partitioned into a 4×4 grid, resulting in 16 patches, each measuring 64×64 pixels.

A custom patch handler is used to process the images. For each sample, the handler extracts all 16 patches and randomly selects k patches to shuffle, where k=2,4, or 6 corresponds to easy, medium, and difficult difficulty levels, respectively. The selected patches are shuffled such that none remain in their original positions, and the shuffled grid is reassembled into a new image. The original, unshuffled image is preserved as the target for supervision.

This setup allows the model to learn to reverse the permutation and reconstruct the spatial layout implicitly. You can see in Fig. 2, above. Refer to the convolution Visual illustration of the self-supervised patch puzzle task at varying difficulty levels. Clockwise from top-left: Original image - the ground truth used for reconstruction supervision. Easy level (k=2) - two patches are shuffled, maintaining most of the spatial context. Medium level (k=4)- four patches are displaced, increasing spatial ambiguity. Hard level (k=6) six patches are permuted with none in their original position, significantly disrupting the visual layout.

Fig. 2. Examples of input-output pairs used in training at varying difficulty levels. The original image (top-left) serves as the target of reconstruction. Shuffled inputs with k=2, k=4, and k=6 patches (top-right, bottom-right, and bottom-left, respectively) represent easy, medium, and hard difficulty levels, where no selected patch remains in its original location.

C. Model Architecture

The proposed model employs a convolutional encoder-decoder architecture specifically designed to reconstruct spatially disordered images generated by patch-level shuffling. This architecture aims to learn both local texture details and global spatial coherence in order to recover the original image layout.

The encoder is composed of a sequence of convolutional blocks. Each block includes a 2D convolution layer followed by batch normalization and a ReLU non-linearity. Stride'd convolutions are used for down sampling, reducing the spatial dimensions while increasing the depth of the feature maps. This progressive compression enables the network to capture increasingly abstract and hierarchical representations of the input image. The encoder transforms the input tensor of size 3×256×256 into a compact latent representation by the end of the stack, effectively summarizing the semantic content and spatial structure.

The decoder is structurally symmetric to the encoder. It consists of transposed convolutional layers (also referred to as deconvolutions) interleaved with non-linear activations, which progressively up samples the latent representation back to the original image resolution. The decoder reconstructs the image by combining global context learned in the encoder with fine-grained spatial details. The final output is passed through a sigmoid activation function to produce normalized pixel values in the range [0, 1], suitable for RGB image reconstruction.

To ensure modularity and facilitate experimentation, the architecture is implemented using reusable blocks for convolution, transposed convolution, and normalization. These building blocks can be easily extended with additional components such as residual connections, dropout layers, or attention mechanisms in future iterations. Overall, this encoder–decoder setup is well-suited for the jigsaw reconstruction task. The encoder learns compact and meaningful representations from disordered input patches, while the decoder is responsible for decoding these features into perceptually coherent and spatially accurate reconstructions. The simplicity of the architecture allows efficient training while maintaining flexibility for potential architectural enhancements.

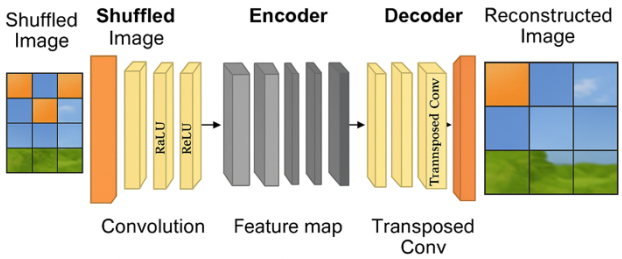

Fig. 3. Convolution encoder-decoder architecture

Fig 3. Shows an overview of the convolutional encoder–decoder architecture. The encoder compresses the shuffled input patches into a latent representation using stacked convolutional layers with down sampling, while the decoder reconstructs the image via transposed convolutions, restoring the original spatial configuration.

D. Loss Formulation

The model is trained using a composite loss function that encourages both pixel-level accuracy and spatial coherence. The primary component is the reconstruction loss, computed using mean squared error (MSE) between the predicted and original images. To enhance the model's ability to infer patch arrangement, we introduce a spatial loss term based on patch-wise cosine similarity. This term measures directional alignment between corresponding patches in the predicted and ground truth images, promoting consistency in feature orientation. The overall objective function is defined as:

$$L_{total} = L_{MSE} + \lambda \cdot L_{spatial} \quad (1)$$

Where (1) is a weighting coefficient that balances the spatial consistency term relative to the reconstruction error.

E. Training Strategy

The network is trained on a subset of 10,000 images from the Places365 dataset. Each image is preprocessed to produce a shuffled version as input and its original counterpart as the reconstruction target. We employ the Adam optimizer with a learning rate of 1×10^{-4} , a batch size of 8, and train the model for 50 epochs using a CUDA-enabled GPU when available. To assess generalization under varying complexity, each training run is conducted using a fixed difficulty level

corresponding to the number of shuffled patches. The reconstructed outputs

RESULTS

F. Evaluation Metrics

To evaluate the quality of reconstructed images in a meaningful and comprehensive manner, we adopt two standard image quality assessment metrics: Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR). These metrics serve complementary roles by capturing both perceptual and pixel-level fidelity.

SSIM measures the structural resemblance between a reconstructed image and its corresponding ground truth by analyzing local patterns of luminance, contrast, and texture. Unlike traditional pixel-wise error metrics, SSIM aligns more closely with human visual perception, making it particularly relevant for tasks like patch-based reconstruction where maintaining structural coherence is critical. The index ranges from 0 to 1, with values approaching 1 indicating higher perceptual similarity.

On the other hand, PSNR quantifies reconstruction accuracy using a signal fidelity perspective, relying on the Mean Squared Error (MSE) between the predicted and ground truth images. Expressed in decibels (dB), PSNR offers a straightforward interpretation of pixel-level differences, with higher values indicating greater similarity. However, because it is sensitive to minor pixel fluctuations and does not reflect perceptual nuances, it is best used in tandem with SSIM for a more balanced evaluation.

To supplement these scalar metrics, we also incorporate patch-wise error visualization maps, which provide spatial insight into where reconstruction failures occur. These visualizations help pinpoint localized inconsistencies, such as misaligned or poorly blended patches, offering valuable diagnostic cues for improving model performance.

G. Evaluation Results

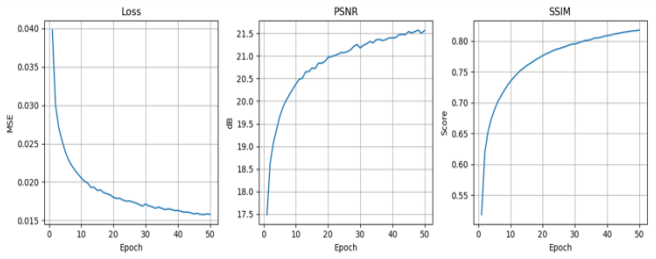

Fig. 4. Training performance curves for MSE, PSNR and SSIM

Fig 4. Shows the training performance curves over 50 epochs. The left plot shows a steady decrease in reconstruction loss (MSE), indicating improved model optimization. The middle plot displays the rising Peak Signal-to-Noise Ratio (PSNR), reflecting enhanced pixel-level fidelity. The right plot presents the Structural Similarity Index (SSIM), highlighting progressive gains in perceptual and structural reconstruction quality. The leftmost plot shows the mean squared error loss, which consistently decreases as training progresses. This trend indicates successful minimization of reconstruction error and suggests that the model effectively learns to reassemble shuffled patches.

The center plot illustrates the PSNR trajectory over the same training period. The values increase steadily from approximately 17.5 dB to over 21.5 dB by the 50th epoch. This upward trend reflects progressive improvements in the pixel-level fidelity of reconstructed images.

The rightmost plot displays the progression of SSIM, which climbs from an initial score near 0.53 to above 0.82 by the end of training. The steep increase during the early epochs highlights the model's rapid acquisition of structural alignment capabilities. The curve then gradually flattens as the model continues to refine finer details in the image structure.

Together, the decreasing loss and increasing PSNR and SSIM curves demonstrate that the network not only improves in minimizing raw error but also becomes more adept at preserving perceptual and structural aspects of the images. These results confirm the effectiveness of the self-supervised reconstruction strategy in learning spatial coherence from shuffled image patches.

In Fig 5. We have the input which is the reshuffled image and the output after the first epoch is blurred out. The second image in Fig 3. Shows the input and output after 50 epochs. We can see significant improvement.

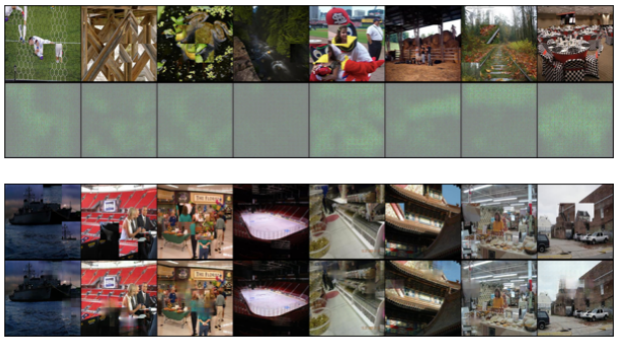

Fig. 5. Input and output after 1 epoch vs 50 epoch

DISCUSSION

The proposed self-supervised framework showcases a promising capability to recover meaningful spatial structure from disordered image patches without relying on any labeled supervision. By casting the task as a permutation recovery problem, the model is encouraged to learn both local pixel-level features and broader spatial dependencies, enabling it to approximate the correct arrangement of shuffled patches with a reasonable degree of accuracy.

Initial qualitative evaluations demonstrate that the model performs reliably under low-difficulty scenarios, where only a small number of patches are rearranged. As the complexity of the task increases i.e., more patches are shuffled, the quality of reconstructions understandably declines. This decline, however, is gradual rather than abrupt, suggesting that the model maintains a degree of spatial awareness even in the face of significant disruption. This observation highlights the framework's robustness in handling varying levels of spatial disorder.

A key contributor to this performance is the inclusion of a spatial consistency term in the loss function. This term not only enforces pixel-wise reconstruction accuracy but also promotes structural alignment across regions, resulting in outputs that are more semantically coherent. Future ablation studies are planned to disentangle the effects of this component and to evaluate its impact relative to other architectural and training choices.

This framework also opens several promising avenues for future research. Incorporating positional encodings or attention-based mechanisms may further enhance the model's ability to reason about spatial relationships across distant patches. Additionally, extending this permutation-based paradigm to video data by applying it to sequences of frames could support the development of models capable of joint spatial and temporal reasoning.

In summary, the system demonstrates strong potential for applications that require a structure-aware understanding of visual content, particularly in scenarios where labeled data is scarce or unavailable. Its efficient and scalable training pipeline, combined with its capacity to learn meaningful representations from self-supervised signals, positions it as a valuable tool in the broader landscape of unsupervised and semi-supervised learning in computer vision.

CONCLUSION

In this work, we proposed a self-supervised deep learning approach for solving visual jigsaw puzzles by reconstructing images from spatially disordered patches. Our method employs a convolutional encoder–decoder architecture that learns to reassemble shuffled image segments by capturing both fine-grained texture information and broader spatial relationships. Performance metrics such as Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR) confirm the model's ability to preserve visual fidelity, particularly in low-complexity settings.

A key component of our framework is the incorporation of a spatial consistency loss, which encourages the network to maintain structural coherence during reconstruction. This results in more perceptually accurate outputs that reflect the original image layout. While the current study limits its scope to puzzles of low difficulty, where only a small number of patches are displaced. The observed outcomes lay a strong foundation for future exploration.

Looking ahead, we plan to extend this framework to more demanding puzzle configurations involving a larger number of shuffled patches. We also intend to investigate more advanced strategies such as superpixel-based segmentation, which introduces irregular yet semantically meaningful regions, and attention-based architectures capable of modeling long-range dependencies between image parts. These enhancements aim to improve the model's scalability and adaptability.

Ultimately, this line of research holds promise for a range of real-world applications, including digital restoration of damaged artifacts, forensic image reconstruction, and scene completion in cases of occlusion or partial visibility. As such, the proposed method not only advances the field of self-supervised representation learning but also opens new avenues for practical deployment in visually complex domains.

FUTURE WORK

This study presents a foundational approach for solving image-based jigsaw puzzles using a deep learning encoder–decoder model, focusing on an easier setting where only a limited number of patches are shuffled. While the results establish a solid proof of concept, several avenues remain unexplored and offer rich potential for future work.

One natural next step is to extend the framework to handle more challenging puzzle configurations. Specifically, medium and high difficulty levels where a greater number of patches are shuffled. Tackling these complex scenarios would allow for a deeper assessment of the model's spatial reasoning and test its ability to cope with increased structural disarray.

Another compelling direction lies in moving beyond the conventional use of uniformly shaped square patches. Future iterations of this work could explore the use of superpixels-irregular, perceptually consistent segments that often align with object boundaries. This shift would allow the model to interpret and reconstruct images at a granularity that is more aligned with human visual perception, potentially improving performance in scenes that contain intricate textures or organic shapes.

Architecturally, enhancements to the current encoder–decoder model could further boost performance. For instance, introducing spatial priors or positional encodings could help the model internalize the original layout of the patches, while attention-based mechanisms such as vision transformers might improve its ability to recognize and relate distant regions of the image especially valuable in more shuffled or complex configurations.

In terms of training strategy, adopting a multi-task learning approach could be beneficial. Combining puzzle solving with related self-supervised tasks like inpainting or contrastive learning might lead to richer and more generalizable visual representations. This could be especially advantageous in scenarios where annotated data is limited or unavailable.

Lastly, the model's utility could extend well beyond academic experimentation. Practical applications include reassembling broken artifacts in archaeology, reconstructing damaged manuscripts in digital humanities, or piecing together anatomical structures from fragmented scans in medical imaging. Each of these domains introduces its own challenges, underscoring the value of adaptable, intelligent puzzle-solving systems.

REFERENCES

[1] Cho, T. S., Avidan, S., Freeman, W. T. (2010). A Probabilistic Image Jigsaw Puzzle Solver. IEEE Transactions on Pattern Analysis and Machine Intelligence

[2] Doersch, C., Gupta, A., & Efros, A. A. (2015). Unsupervised Visual Representation Learning by

Context Prediction. Proceedings of the IEEE International Conference on Computer Vision (ICCV).

[3] Noroozi, M., & Favaro, P. (2016). Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[4] JigsawGAN: Auxiliary Learning for Solving Jigsaw Puzzles with Generative Adversarial Networks (Lee et al., 2020)

[5] Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image Quality Assessment: From Error Visibility to Structural Similarity. IEEE Transactions on Image Processing

[6] D. Bridger, D. Danon, and A. Tal, "Solving Jigsaw Puzzles With Eroded Boundaries," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3526–3535, 2020. doi

[7] A. Yu and K. Römer, "Solving Jigsaw Puzzles with Linear Programming," in Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2016, pp. 48–64.

[8] T. Paumard, K. D. Kania, S. O. Ba, J. Pinquier, and J. L. Rouas, "Solving Jigsaw Puzzles with Combined Deep and Graphical Models," arXiv preprint arXiv:2008.12959, 2020.

[9] C. Saharia, J. Ho, T. Salimans, W. Chan, D. J. Fleet, and M. Norouzi, "ImageBART: Bidirectional Context with Multinomial Diffusion for Autoregressive Image Generation," arXiv preprint arXiv:2206.00790, 2022.