Image Super-Sampling and Reconstruction from

Sparse Grid Samples

Yiming Dou, Qing Yang

**Abstract**:

**Key word：Voronoi-Delaunay Triangulation, UNet, Masked-Auto-Encoder, Image reconstruction**

# Introduction

## Significance

Real-time rendering application is ubiquitous in modern media such as mobile phone and virtual reality. With increasing display and resolution demands, low processing latency and low resources used pose a challenge for today’s rendering philosophy. Fixed foveated rendering renders peripheral regions at low resolutions. Microsoft reduces shading complexity for foveated and high-resolution displays using variable rate shading. However, these methods rely heavily on specific hardware; or brings some artificial features in image which lower the display quality.

In this paper, two methods of sparse-sampling are proposed, including a random sampling algorithm based on Delaunay Triangulation and a patch-sampling algorithm based on FFT high-pass filter. After the sampling, the image is stored as sparse grids instead of traditional dense matrix. In order to reconstruct the original image from the sparse grids, two image reconstruction algorithms are designed accordingly, including image super-sampling based on a network that is similar to UNet and a image reconstruction algorithm based on Masked Auto Encoder (MAE). To summarize, the contribution of this article is threefold:

1. Introducing a novel way of storing and representing images as sparse grids instead of dense grids
2. Proposing a sparse-grid sampling method based on Delaunay Triangulation and accordingly design a model to reconstruct the original image based on the sparse grids.
3. Proposing a sparse-patch sampling method based on image patch splitting and FFT and design a model to reconstruct the original image based on the chosen image patches.

## Article Structure

As is shown in the following figure, the article architecture consists of **two independent paths** that respectively proposes an image sampling and reconstruction system. In the first path, a sparse-grid sampling method is designed upon Delaunay Triangulation and a model based on UNet is designed to reconstruct the original image. In the second path, a sparse-patch sampling method is designed using FFT and patch splitting, while the image reconstruction algorithm is accordingly designed with the help of Masked Auto Encoder (MAE).

**Each** of the paths introduce a novel algorithm that sample the original image to sparse grids and accordingly design an image super-sampling or reconstruction algorithm.

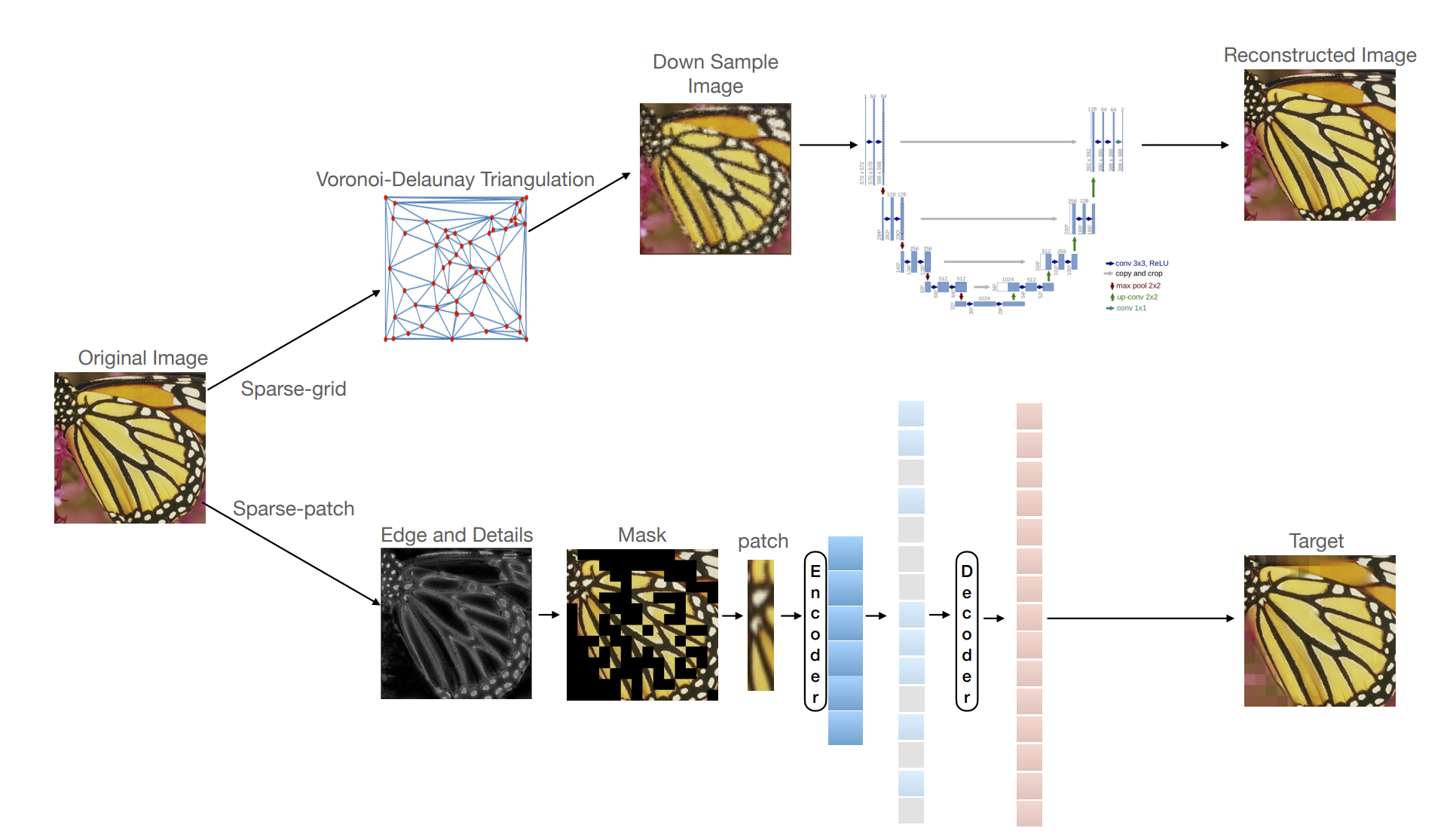


Fig 1: Workflow of this article

# Related Works

## Image Supersampling

Rendering the image at a much higher rsolution that the one being displayed and then calculating an average color value, supersampling removes aliasing to make realistic objects. Algorithms include Grid algorithm in uniform distribution, Rotated grid algorithm, random algorithm, Jitter algorithm, Posson disc algorithm, Quasi-Monte Carlo method algorithm, N-books, RGSS, HRAA, Filpquad, Fliptri have been proposed to against alias. However, basic supersamling methods are computational expensive due to the demand for video card memory and memory bandwidth. To solve this problem, machine learning or so called deep learning has been introduced to this field. Nvidia proposed deep learned supersamling(DLSS), exploiting the neural network to upsampling low resolution content. [28] 4x4 supersamling with high spatial and temporal fidelity has been achieved using temporal information and representation learning framework [29]. In our work, we also incorporate machine learning into the supersampling technique to reach a better display quality and lower computation cost.

## Real-time Rendering

Real-time rendering focuses on analyzing and producing images in real time. It’s widely used in today’s computer, mobile, virtual reality and so on. Users can interact with the render as it is developed. Besides, real-time software rendering facilitate its application. However, demands for real-time requires large amount of computation resources; and specific hardware in some application.

## Fast Fourier Transform

Fast Fourier transform (FFT) is an algorithm that widely implemented in programming language package and compatible with GPU. FFT sampling can convert signal including the image to the spectrum domain. It’s widely used in designing digital filter and fast processing of image because it can expose image features, like periodic interferences which is not visible in spatial domain and compress image into a more compact representation.

[26] evaluates the performances of FFT in real-time application and suitability for GPU implementation. Besides, conditions that FFT gives better performance have been identified. FFT is utilized to extract a sparse and collaborative for image classification [27]. In our discussion, FFT is used as a sampling method that lower the resolution of the original image thus the resource needed for following tasks.

## Disentangled Representation Learning

Disentangled representation is desired as it represents a human interpretable pattern [1,2,3], enabling the downstream tasks learned more easily [4] and generalizes better [5]. In this paper, we try to use the disentangled property of represesntation learning to extract a disentangled latent space which is used as the input for the reconstruction.

We notice the recent study of disentanglement is promoted by two communities: Disentanglement in Deep Features and Independent Component Analysis. Their research previously lie on different assumptions, data patterns, and evaluation metrics.

One community is motivated by the newly raised deep learning for encouraging disentangled representation over independent factors. they have shown much empirical progress on this problem and they directly term their goal as ``disentanglement''. The related study is usually based on deep generative models. For instance, VAE-based methods have achieved successes on this task [6,7,8,9]. Besides, Generative Adversarial Networks (GAN) [2,10] are also put into the discussion of encouraging representations' disentanglement. More recently, people have shown that the GAN-based approach can achieve competitive performance as the above VAE variants [11,12,13]. A recent work [14] summarizes the popular methods and metrics in this community and proposes a tool for evaluation called disentanglement\_lib, including popular metrics such as DCI [15], SAP [16], MIG [8] and so on. We use the encoder with verified high disentanglement score in the sampling.

Besides this series of studies, exploring underlying factors of variation in data pattern is a long-standing goal of the Independent Component Analysis (ICA) community [17]. They share many similarities, for example, generative models, e.g., VAEs, are recently popular in both [18,19]. ICA usually has different assumptions with the ``purely unsupervised learning'' [20]. For example, the pattern of noise [21,22] or some additional auxiliary variables [23,24] can be observed. Traditionally, ICA uses identifiability to assess their desired representation pattern and the popular metric is Mean Correlation Coefficient (MCC). SlowVAE [25] recently makes a great effort to connect the two branches of study but it still requires additional information such as temporal transition pattern.

# Contents and Methods(or Algorithm)

In this article, two sparse-grid down-sampling methods are proposed, and two methods that reconstruct the origin image from sparse grids are designed accordingly.

## Sparse-Grid Down-sampling

Traditional down-sampling methods regard the image as a matrix with several channels, and the sampling points are covered on the original image densely and uniformly, which results in the low-resolution (LR) image being dense-and-uniform grids. This kind of methods may lead to several problems:

1. Data Redundancy: The position of sampling points are uniformly and densely generated, thus not being able to distinguish the detailed features in the image. Specifically, it is not reasonable to sample the areas with more details and those with less details from the same sampling-rate, since this may cause data-redundancy. If the same amount of data is used, and the data sampled from areas with more detail occupies a greater proportion, then more details of the image are retained, resulting in better effect.
2. Computation Redundancy: In addition to data redundancy, the computation redundancy is also a serious problem. Consider an image rendered from a model, much computation such as lighting and shading is needed, and these computations usually result in heavy cost. If the shading and lighting procedures of two sample points are similar with each other, then there is no need to sample both of them.

In order to solve these possible problems, two algorithms are proposed in this article, which use sparse sample points to down-sample the image.

### Voronoi-Delaunay Random Point Triangulation Algorithm

In mathematics and computational geometry, a Voronoi-Delaunay Triangulation can be depicted as: Given a set P of discrete points in a general position, DT(P) is a triangulation such that no point in P is inside the circumcircle of any triangle in DT(P). Delaunay Triangulation maximize the minimum angle of all the angles of the triangles in DT(P).

In order to efficiently compute the triangulation process, the Divide-and-Conquer method is used, which has been shown to be the fastest method to compute Delaunay Triangulation. In this algorithm, a line that split the vertices into two non-overlap subsets is drawn recursively. Next, the Delaunay triangulation of each subset is computed, after which the two subsets are merged along the splitting line. The merge operation can be done within time O(n), thus the total running time is O(nlogn), which is acceptable.

In this article, 2D Voronoi-Delaunay Triangulation is used to generate the sample points on the images. Given the number of the sample points, the position of the points are randomly generated within the range of . Then, the points are scaled to fit the size of the image and covered on the image, which is shown in the following figure.

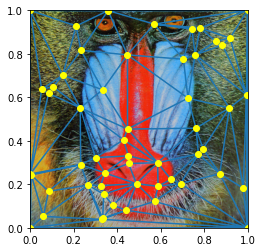


Fig 2: Sample points generated by Voronoi-Delaunay Triangulation

Next, the Delaunay Triangulation algorithm is performed, which generates triangles from the randomly-generated sample points. This results in each pixel’s staying in one and only one triangle, making it possible to set the color of each pixel based on the colors of the vertices of the triangle. Specifically, the color of each pixel of the LR image is set as the average color value of the three vertices. The reason of this is that it requires relatively less computation, and the generated LR image is shown in the following figure.



Fig 3: LR image generated by sparse-sampling

### Image Patch Splitting

This method follows the procedure of Vision Transformer (ViT). Each image is split into fixed-size patches. In the next step, the patches that contain more detailed information should be sampled, since these patches bring more information into the model and tend to help it reconstruct the original image better.

To find out these patches, a high-pass filter based on FFT is used to extract the texture and edge features of the image. The process can be depicted in the following figure:



Fig 4: High-pass filtering based on FFT

After the high-pass filtering, the areas with more details are retained, making it possible to select the patches that provides more information. Specifically, to show the effect of selecting patches based on FFT high-pass filtering, we adopt two selection strategy: (1) Random-Selection, (2) FFT-based-Selection. The result of each strategy is shown in the following figure (the patches that are not selected are depicted as masks):

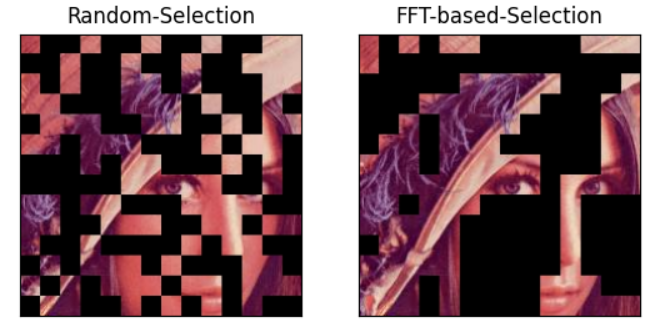


Fig 5: Comparison between two selection strategies

Therefore, the FFT-based selection is able to obtain the patches that contains more information (such as those representing hair and eyes). Moreover, during the computation and data storage, the masked patches are not needed, the selected patches are stored as sparse-grid samples, thus the computation and data redundancy can be largely reduced.

## Image Super-Sampling and Reconstruction

Image Super-Sampling algorithms take the LR image as the input and reconstruct the HR image. In this article, for each down-sampling algorithm, one image reconstruction method is proposed accordingly. Therefore, **2 methods** of reconstruction methods are proposed in this project.

### Image Super-Sampling based on UNet

In order to reconstruct the HR image from the LR image sampled by Voronoi-Delaunay Triangulation, a network that is similar to the structure of UNet is used.

UNet, which is firstly proposed in 2015 for image segmentation, is build upon is so-called “fully connected convolutional network”. The idea is supplementing the basic contracting network by adding successive layers, in which pooling layers are replaced by up-sampling layers. Therefore, these layers are able to increase the output resolution. Combining the information from the contracting and expansive layers, the network performs better at learning pixel-wise tasks such as segmentation or super-sampling.

T During the contraction, the spatial information is reduced while feature information is increased. The expansive pathway combines the feature and spatial information through a sequence of up-convolutions and concatenations with high-resolution features from the contracting path. A large number of up-sampling parts allow it to propagate context information to higher resolution feature layers.

The structure of the network is shown in the following figure. The network consists of a contracting path (left side) and an expansive path (right side), which gives it the U-shaped architecture. The contracting path is a typical convolutional network consisting of convolution layers that extract features, max-pooling layers that down-sample the feature and ReLU activation layers. The expansive path includes up-sampling layers followed by convolution layer (up-convolution), which increases the resolution of the input. The final layer is a simple 1x1 convolutional layer. In all, the whole network includes 23 convolutional layers.

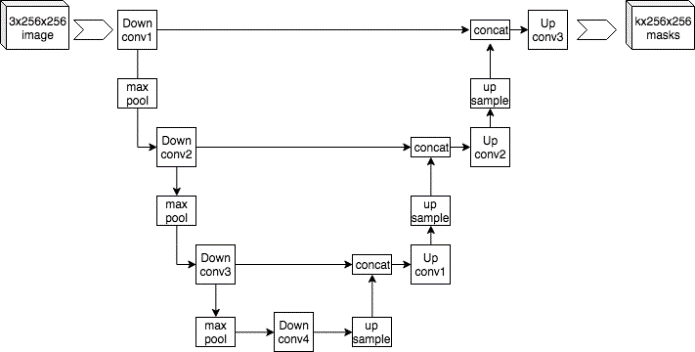


Fig 6 Structure of UNet

The LR image is input into the network and the output is the reconstructed image. In order to minimize the difference between the reconstructed image and the origin image, the Mean-Squared-Error (MSE) loss is used to meter the effect of the reconstruction. The whole structure is depicted in the following figure:

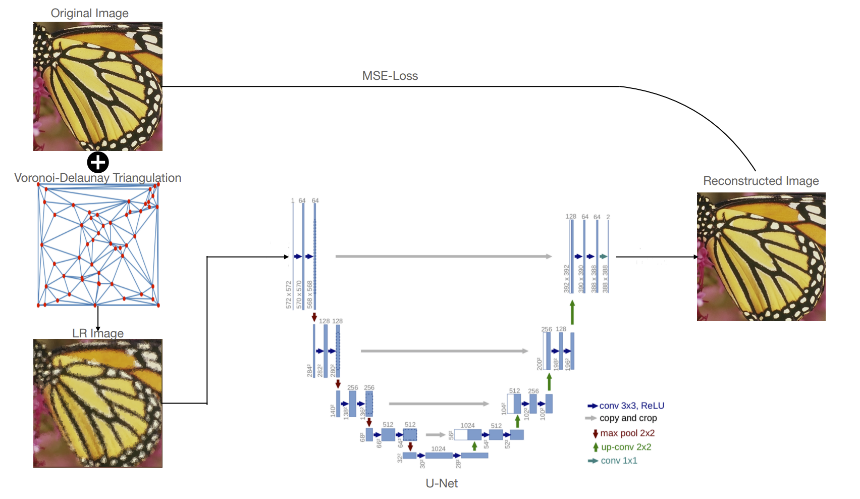


Fig 7: Work-flow of Image super-sampling

### Image Reconstruction based on Masked-Auto-Encoder (MAE)

In order to reconstruct the origin image based on the sparse patches, the following problem must be solved: Since the input is no longer the complete image, the super-sampling procedure should not base on models that are only able to extract local features (such as CNNs), then how should we organize the input?

The transformer-based Masked-Auto-Encoder (MAE) is an excellent choice. The input of this model is patches with any amount, and the output is the whole predicted-image. The MAE is a autoencoder that reconstructs the original image given the partial observation of it. During the encoding process, the patches of the image is sent into the network, generating the latent representation of the image and in the decoding process, the representation is decoded back into image, which is the reconstructed image. The architecture of MAE is shown in the following figure:

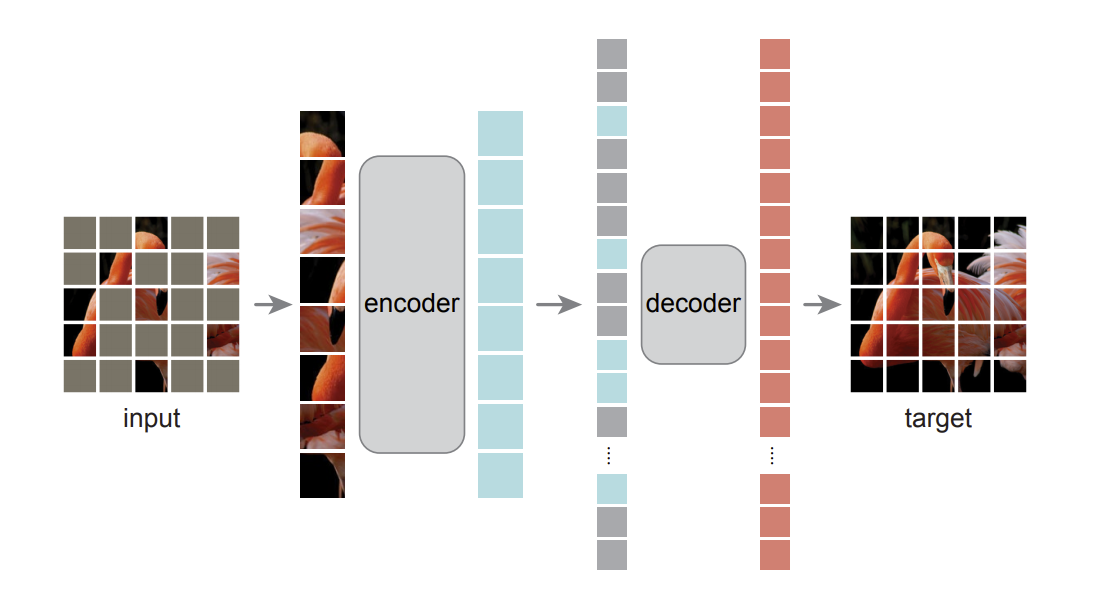


Fig 8: MAE architecture

In detail, the MAE consists of these parts:

1. Masking

Each image is split into non-overlapping patches, which is similar to Vision Transformer (ViT). Next, using the methods proposed above, a subset of the patches are chosen and others are removed. This approach significantly reduces the data and computation redundancy, and the sparse input makes it possible for designing a large and efficient encoder.

1. Encoding

The encoding method is similar to that of ViT, the input patches are firstly embedded by a linear projection layer and are added with positional embeddings. Specially, the MAE only uses the chosen patches, meaning that the masked patches are removed from the input. This allows the model to use a relatively large encoder, compared with traditional methods.

1. Decoding

Different from the encoder, the decoder takes everything into input, including the chosen patches and the masked tokens so as to reconstruct the origin image. In the output, each masked token is a vector learned by the model, predicting the origin image.

1. Reconstruction Target

The MAE generates a pixel-wise reconstruction for each masked patch. Each value in the output of the decoder represents the pixels of a patch. In order to get the reconstruction image, the output of the decoder is reshaped. In order to evaluate the difference between the reconstruction and the original image, the Mean Squared Error (MSE) loss function is used.

Therefore, the value of each pixel of the origin image can be predicted based on sparse-grid patch samples.

In this article, the target image is reconstructed following the procedure shown in the following figure:

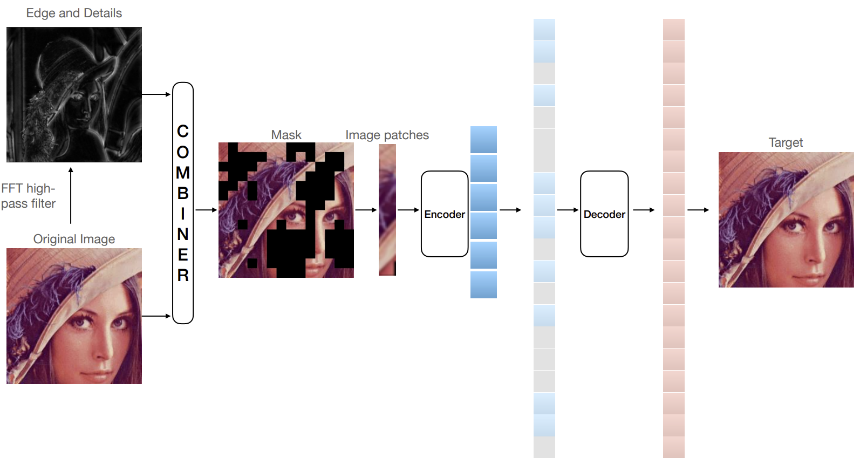


Fig 9: Work-flow of image reconstruction using MAE

Firstly, the original image is input into a FFT high-pass filter, and the edges and details will be retained by the filter.

Secondly, we sparsely sample on the original image by combining the information of the original image and the image details, making sure that the patches with more details are chosen.

Thirdly, the chosen patches are embedded and added with positional embeddings.

Fourthly, the embeddings are sent into the MAE, which extracts the representation of the patches and further predicts the values of the masked patches.

Finally, the predictions are reshaped and the image is reconstructed by combining the chosen patches and the predicted patches.

# Experiment Results and Analysis

## Setup

## Image Super-Sampling based on Voronoi-Delaunay Triangulation Samples

In this section, the experiments of image super-sampling based on the sample points whose amount ranges from 500 to 20000 are conducted. The results of the experiments are shown in the following figure:

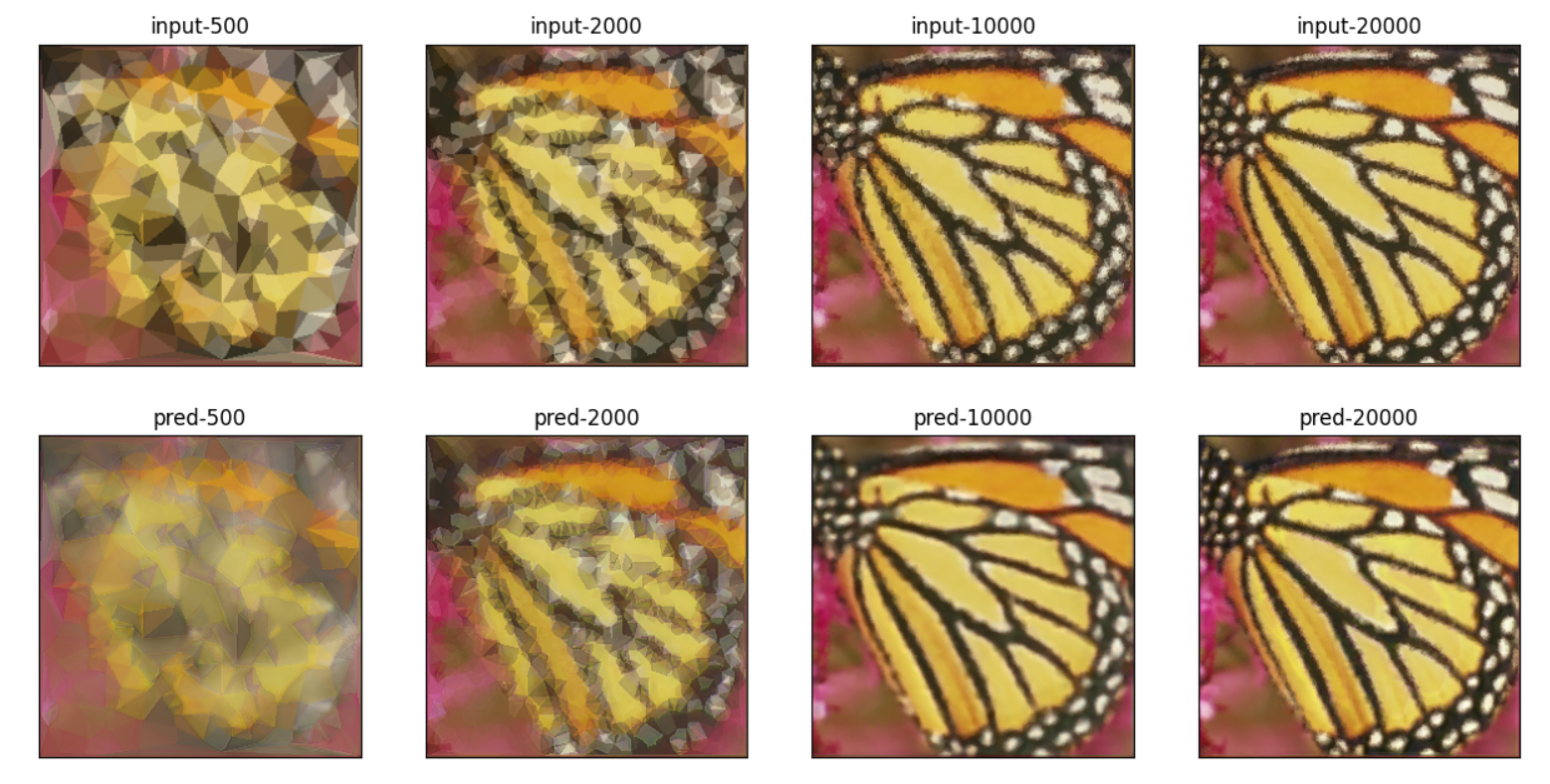


Fig 10: Comparison between different sampling rates

Moreover, the PSNR of each experiment is also computed, which is shown in the following table:

|  |  |
| --- | --- |
| Sample Point Count | PSNR |
| 500 | 13.03 |
| 2000 | 14.58 |
| 10000 | 19.60 |
| 20000 | 20.91 |

## Image Reconstruction based on Masked-Auto-Encoder (MAE)

In this section, the reconstruction results of the patches are depicted. In order to compare the reconstruction results, the mask-ratio ranging from 0.1 to 0.9 are adopted. The results of the experiments are shown in the following figure:

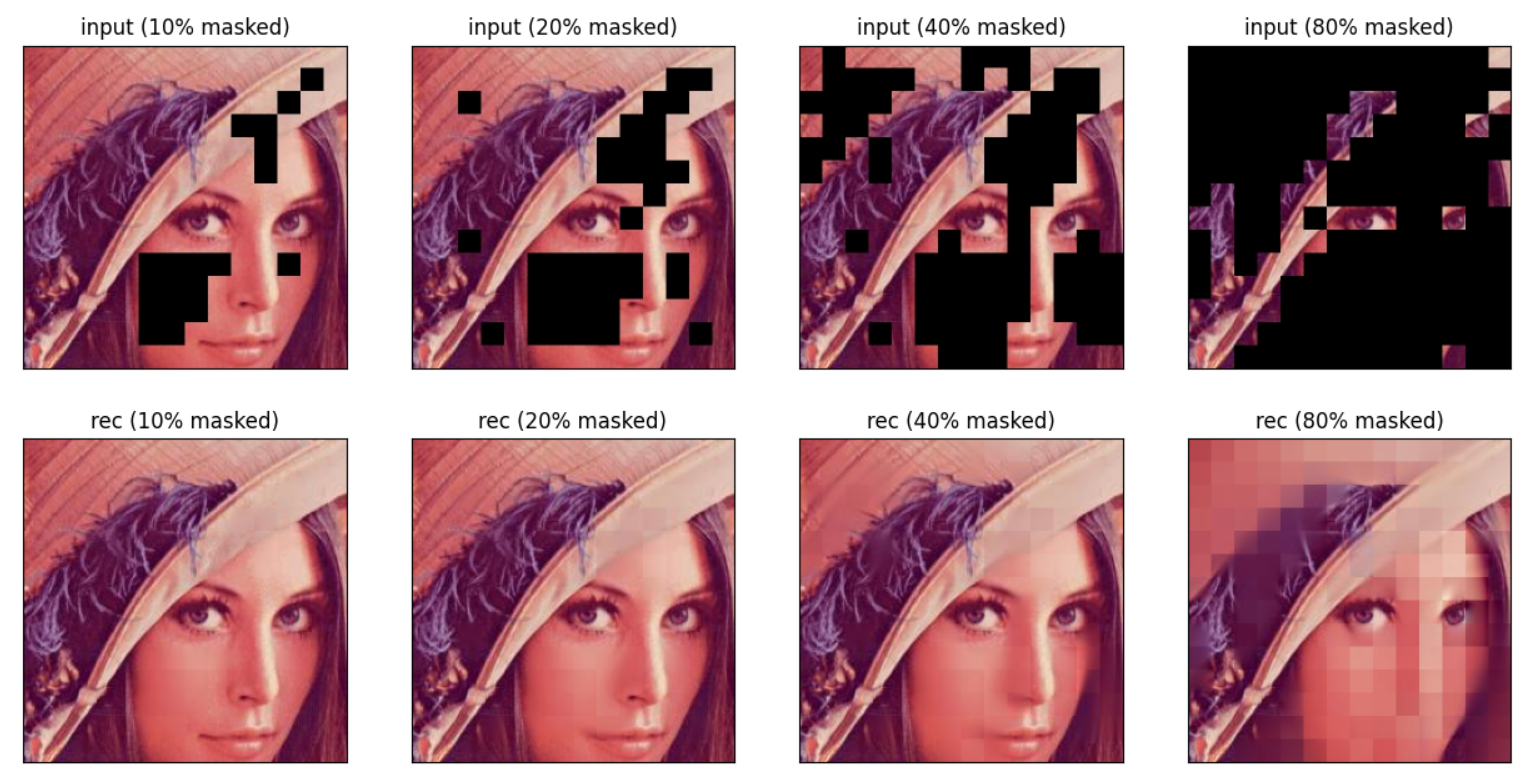


Fig 10: reconstruction effects (FFT-based mask)

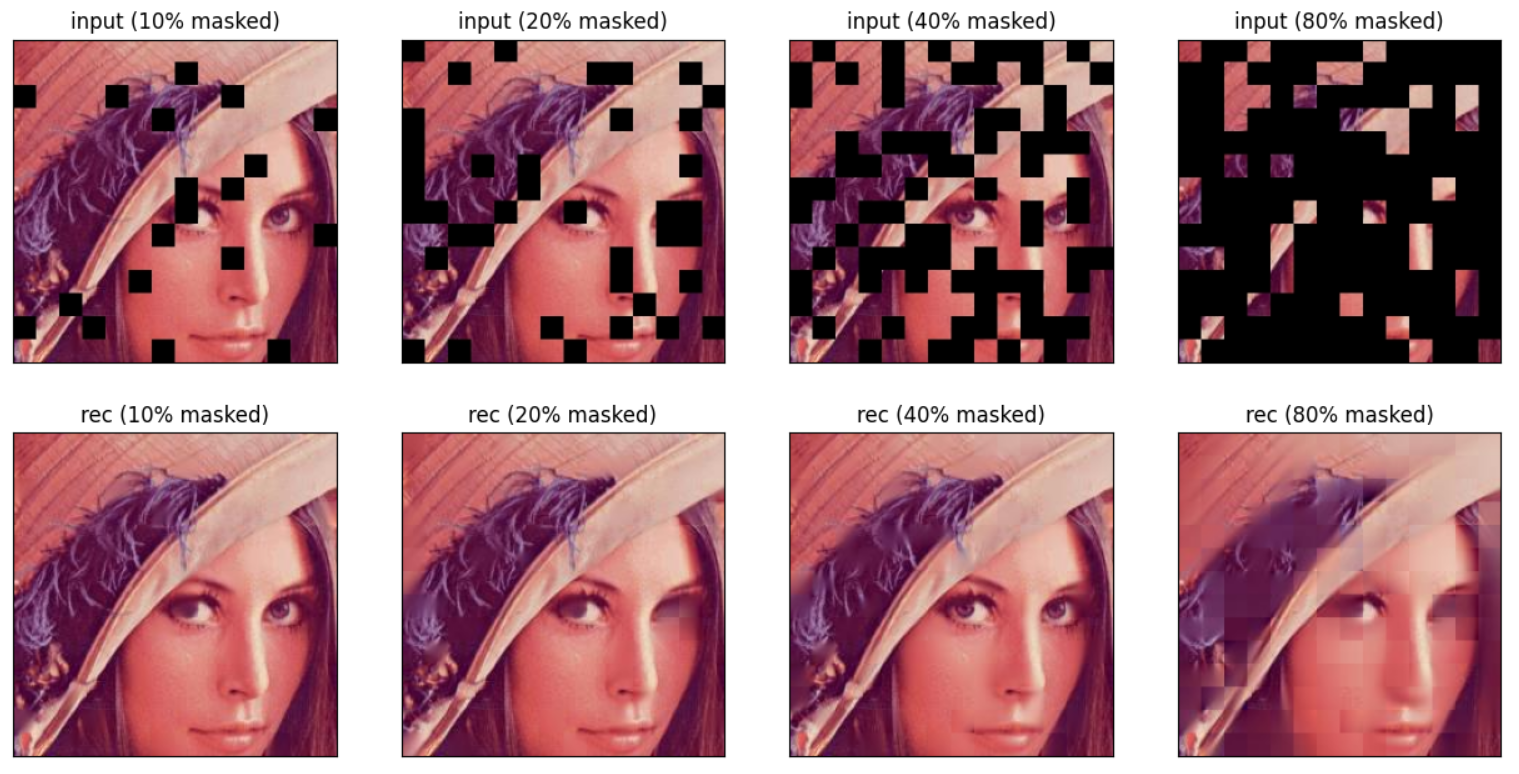


Fig 11: reconstruction effects (random mask)

It is shown by the figure that the random masking strategy may cause very bad reconstruction effect even if masking ratio is low (20%). On the contrary, the FFT-based masking ratio tend to generate better and more stable results, especially when masking ratio is relatively low.

# Distinctive or Innovation Points

# 补充说明：

References:

1. Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.
2. Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 2180–2188, 2016.
3. Tejas D Kulkarni, Will Whitney, Pushmeet Kohli, and Joshua B Tenenbaum. Deep convolutional inverse graphics network. *arXiv preprint arXiv:1503.03167*, 2015.
4. Sjoerd van Steenkiste, Francesco Locatello, Ju ̈rgen Schmidhuber, and Olivier Bachem. Are disen- tangled representations helpful for abstract visual reasoning? *arXiv preprint arXiv:1905.12506*, 2019.
5. Alessandro Achille, Tom Eccles, Loic Matthey, Christopher P Burgess, Nick Watters, Alexander Lerchner, and Irina Higgins. Life-long disentangled representation learning with cross-domain latent homologies. *arXiv preprint arXiv:1808.06508*, 2018.
6. Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *arXiv preprint*, 2016.
7. Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pp. 2649–2658. PMLR, 2018.
8. Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disen- tanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, 2018.
9. Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentan- gled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*, 2017.
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.
11. Zinan Lin, Kiran Thekumparampil, Giulia Fanti, and Sewoong Oh. Infogan-cr and modelcentrality: Self-supervised model training and selection for disentangling gans. In *International Conference on Machine Learning*, pp. 6127–6139. PMLR, 2020.
12. Insu Jeon, Wonkwang Lee, and Gunhee Kim. Ib-gan: Disentangled representation learning with information bottleneck gan. *arXiv preprint*, 2018.
13. Wonkwang Lee, Donggyun Kim, Seunghoon Hong, and Honglak Lee. High-fidelity synthesis with disentangled representation. In *European Conference on Computer Vision*, pp. 157–174. Springer, 2020b.
14. Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Scho ̈lkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pp. 4114–4124. PMLR, 2019.
15. Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disen- tangled representations. In *International Conference on Learning Representations*, 2018.
16. Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentan- gled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*, 2017.
17. Aapo Hyva ̈rinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.
18. Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoen- coders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intel- ligence and Statistics*, pp. 2207–2217. PMLR, 2020a.
19. David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. *arXiv preprint arXiv:2007.10930*, 2020.
20. Hermanni Ha ̈lva ̈, Sylvain Le Corff, Luc Lehe ́ricy, Jonathan So, Yongjie Zhu, Elisabeth Gassiat, and Aapo Hyvarinen. Disentangling identifiable features from noisy data with structured nonlinear ica. *arXiv preprint arXiv:2106.09620*, 2021.
21. Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in Neural Information Processing Systems*, 29:3765–3773, 2016.
22. Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoen- coders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intel- ligence and Statistics*, pp. 2207–2217. PMLR, 2020a.
23. Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 859–868. PMLR, 2019.
24. IlyesKhemakhem,RicardoPioMonti,DiederikPKingma,andAapoHyva ̈rinen. Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ica. *arXiv preprint arXiv:2002.11537*, 2020b.
25. David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. *arXiv preprint arXiv:2007.10930*, 2020.
26. Fialka, Ondirej, and Martin Cadik. "FFT and convolution performance in image filtering on GPU." In Tenth International Conference on Information Visualisation (IV'06), pp. 609-614. IEEE, 2006
27. Tian, Chunwei, Qi Zhang, Guanglu Sun, Zhichao Song, and Siyan Li. "FFT consolidated sparse and collaborative representation for image classification." Arabian Journal for Science and Engineering 43, no. 2 (2018): 741-758.
28. Andrew Edelsten, Paula Jukarainen, and Anjul Patney. 2019. Truly next-gen: Adding deep learning to games and graphics. In NVIDIA Sponsored Sessions (Game Developers Conference)
29. Lei Xiao, Salah Nouri, Matt Chapman, Alexander Fix, Douglas Lanman, and Anton Kaplanyan. 2020. Neural Supersampling for Real-time Rendering. ACMTrans. Graph. 39, 4, Article 142 (July 2020), 12 pages. https://doi.org/10. 1145/3386569.3392376

时间安排与分工统计表

|  |  |  |  |
| --- | --- | --- | --- |
| **组员信息**（含组长） | | | |
| 学生姓名 | 窦铱明 | 学 号 | 519021910366 |
| 项目分工 |  | | |
| 学生姓名 | 杨晴 | 学 号 | 519021910756 |
| 项目分工 |  | | |
| 学生姓名 |  | 学 号 |  |
| 项目分工 |  | | |
| **时间安排/**  **Schedule** | （如选题、方案制定、试验研究、数据处理、研制开发、撰写总结报告等）(Such as topic selection, program formulation, experimental research, data processing, research and development, writing summary reports, etc.) | | |

附GitHub提交记录：

