Sharif University of Technology
Computer Engineering Department

Master's Thesis Defense

Evaluation of Explainability Methods for Breast Cancer Histopathological Image Classification.

Pardis Afshar

Supervisors:
Prof. Emad Fatemizadeh
Prof. Mohammad Hossein Rohban

Examiners:
Prof. Mahdieh Soleymani Baghshah
Prof. Hoda Mohammadzade

Jan 2025

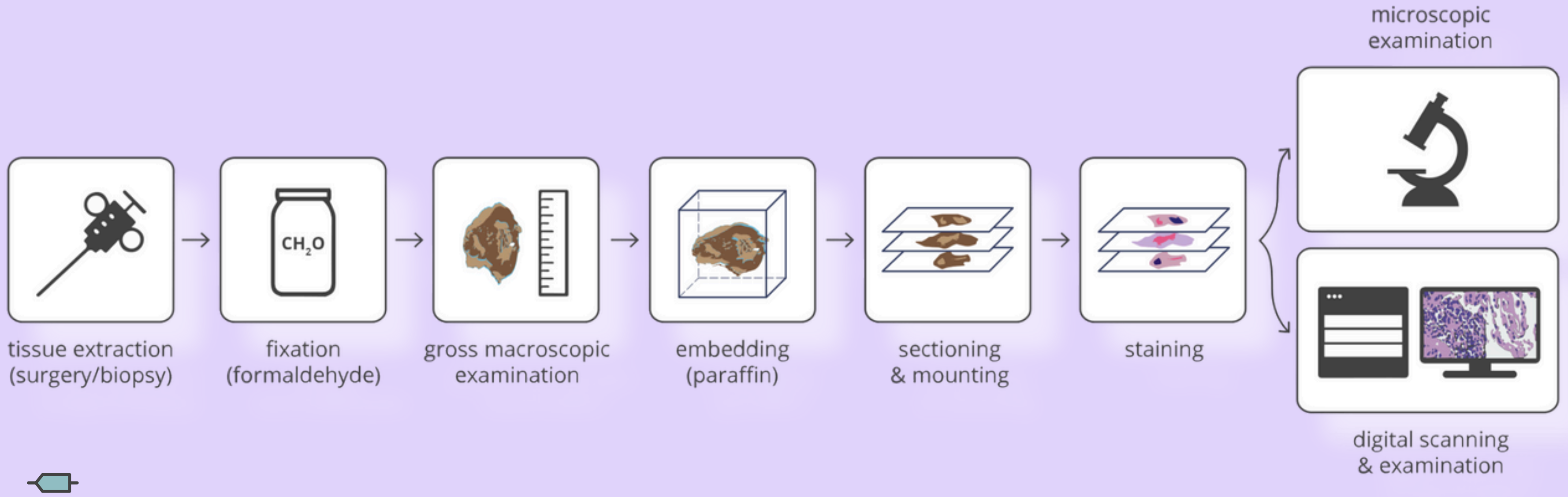
Overview

- Introduction
- Related Works
- Challenges
- Proposed Methodology
- Results
- Conclusion

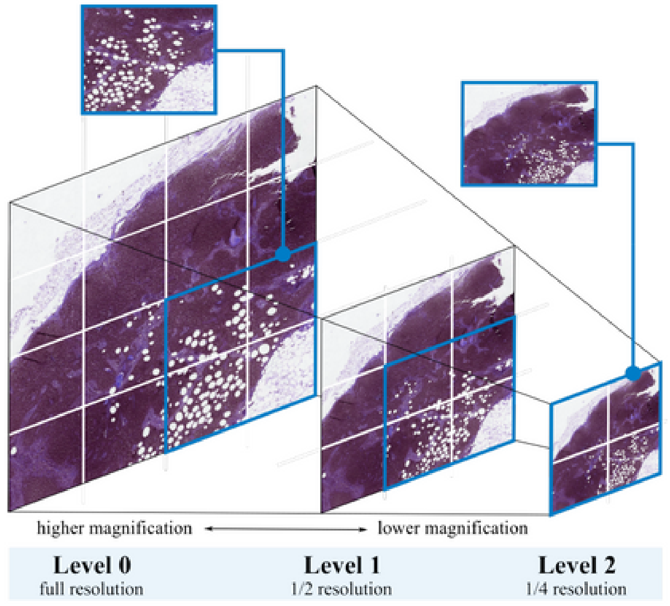
Breast Cancer

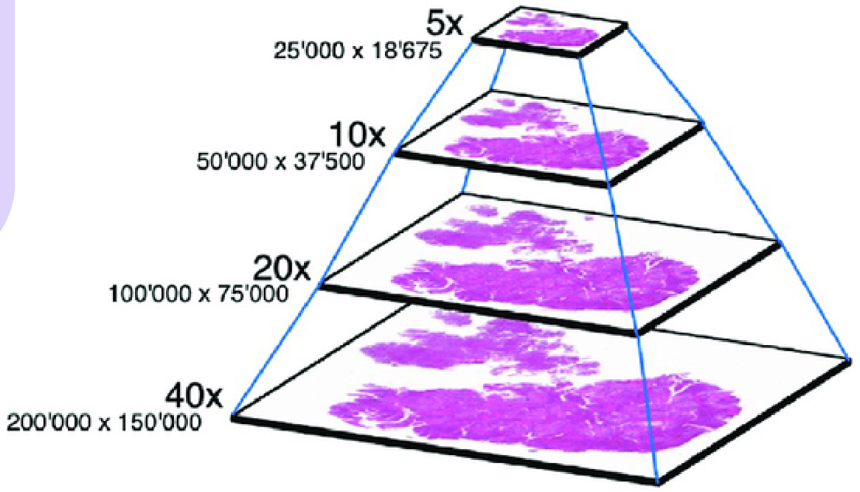
- **Breast Cancer:**
  - One of the most common types of cancer among women
  - Difficult to diagnose
  - Requires early detection

- **Diagnostic Methods:**
  - Mammography
  - MRI (Magnetic Resonance Imaging)
  - Ultrasound
  - Biopsy



The tissue preparation process in pathology includes fixation, tissue processing, and staining. These steps are performed for examining samples under a microscope or scanning and digitizing them for further analysis.





- **Challenges:**

1. Gigapixel Image Sizes
2. Time-Consuming for Pathologists
3. Error-Prone Evaluations
4. Pathologist Fatigue

AI (Machine/Deep Learning)

Black Box Problem

- **Key Characteristics for Gaining the Trust of Doctors and Specialists in an AI Model:**

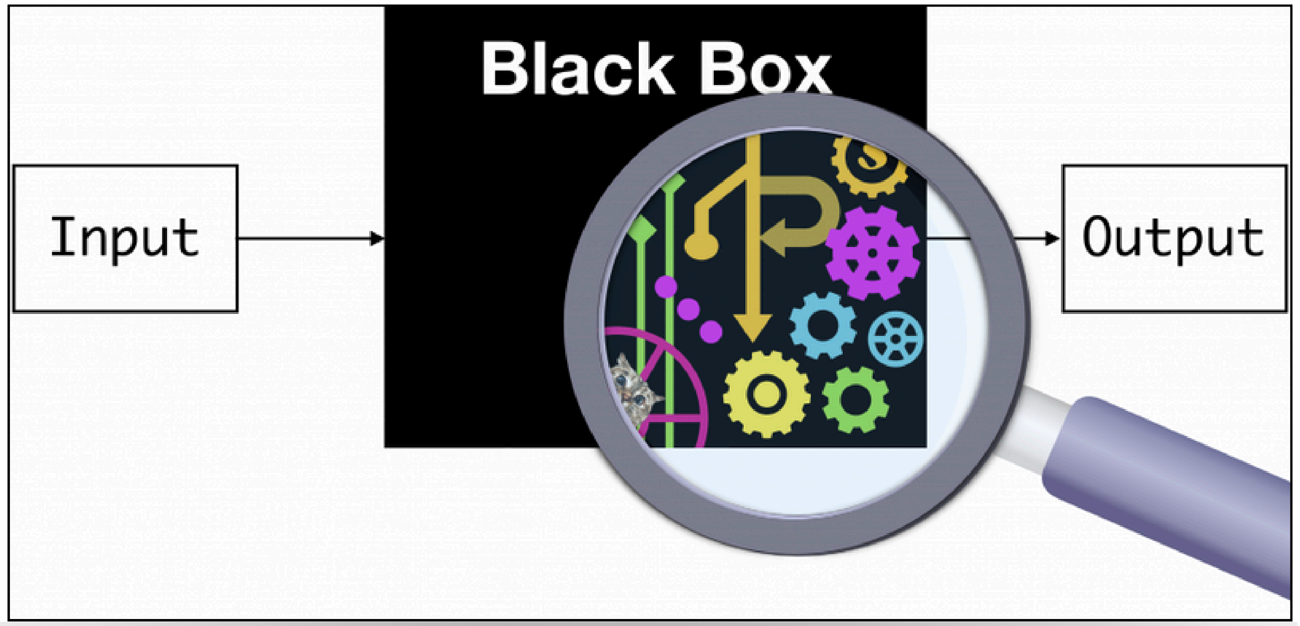
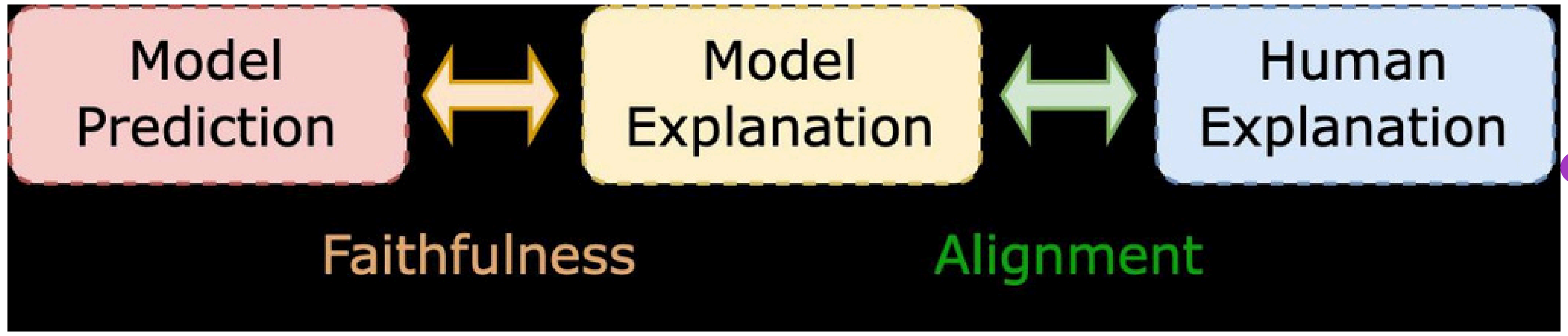


Figure 1: Schematic illustration of the black-box nature of AI models



01 High-Accuracy Classifier Model ✓

02 Interpretability of the Classifier Model ✓

If conditions are met The explanations provided by methods like XAI will be more aligned with the actual decision-making process of the model.

03 Evaluation of Explanations Generated by XAI Methods

"Explanations must faithfully reflect the model's predictions and **align with human reasoning**." ✗

Classification of Breast Cancer Histopathological Images

- **Binary Class**
  - VGG-16
  - ResNet-50
- **Multi Class**
  - VGG-16
  - ResNet-50
  - InceptionNet

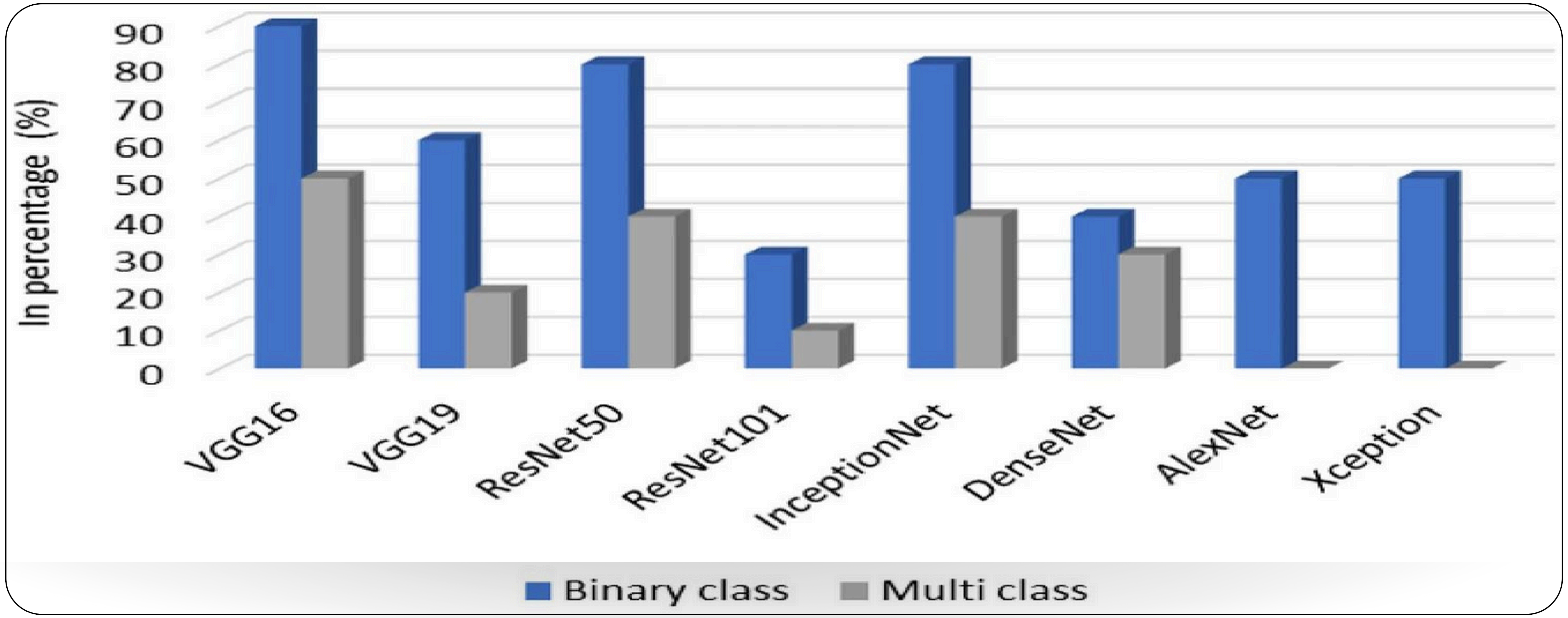


Figure 2: Various CNN Architectures for Binary and Multi-Class Classification, Specifically for Breast Cancer Histopathological Images

Explainable Artificial Intelligence

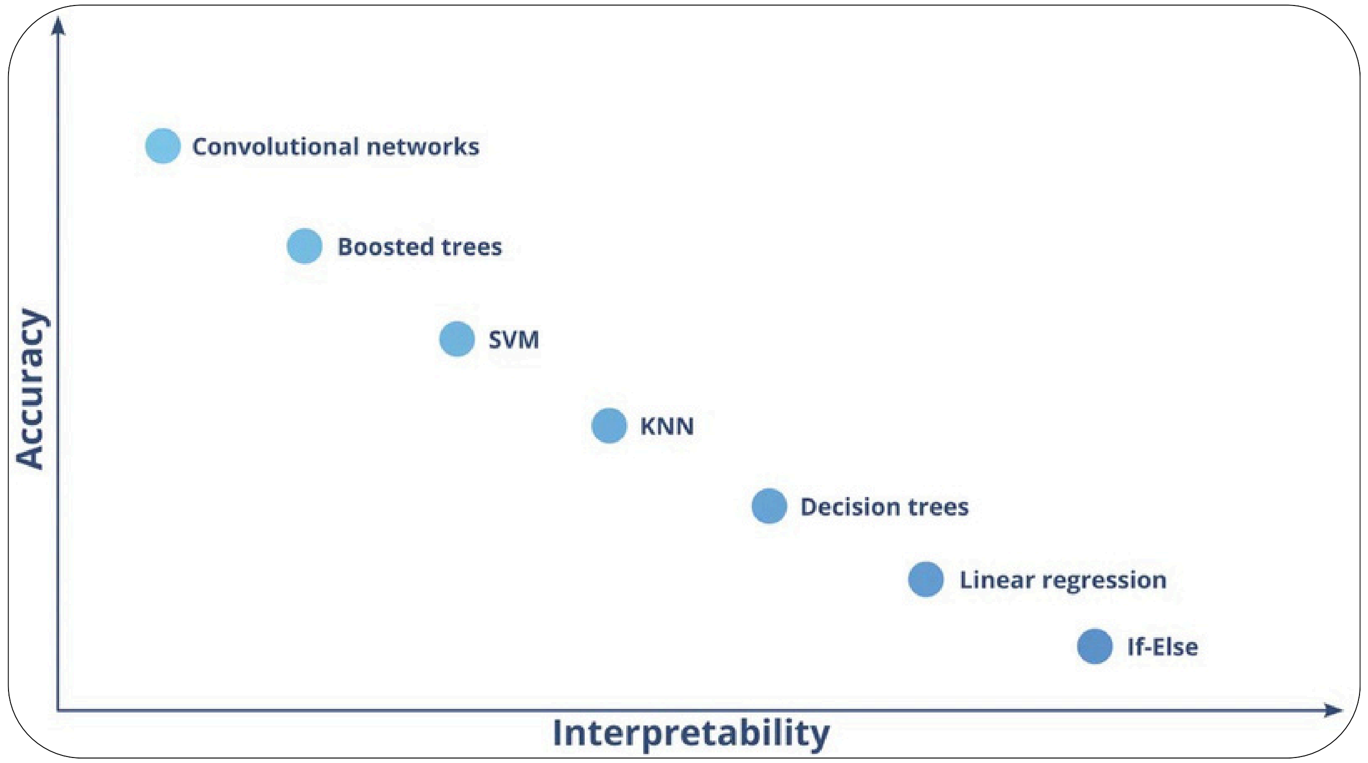


Figure 3: The Relationship Between Accuracy, Complexity, and Interpretability of Models

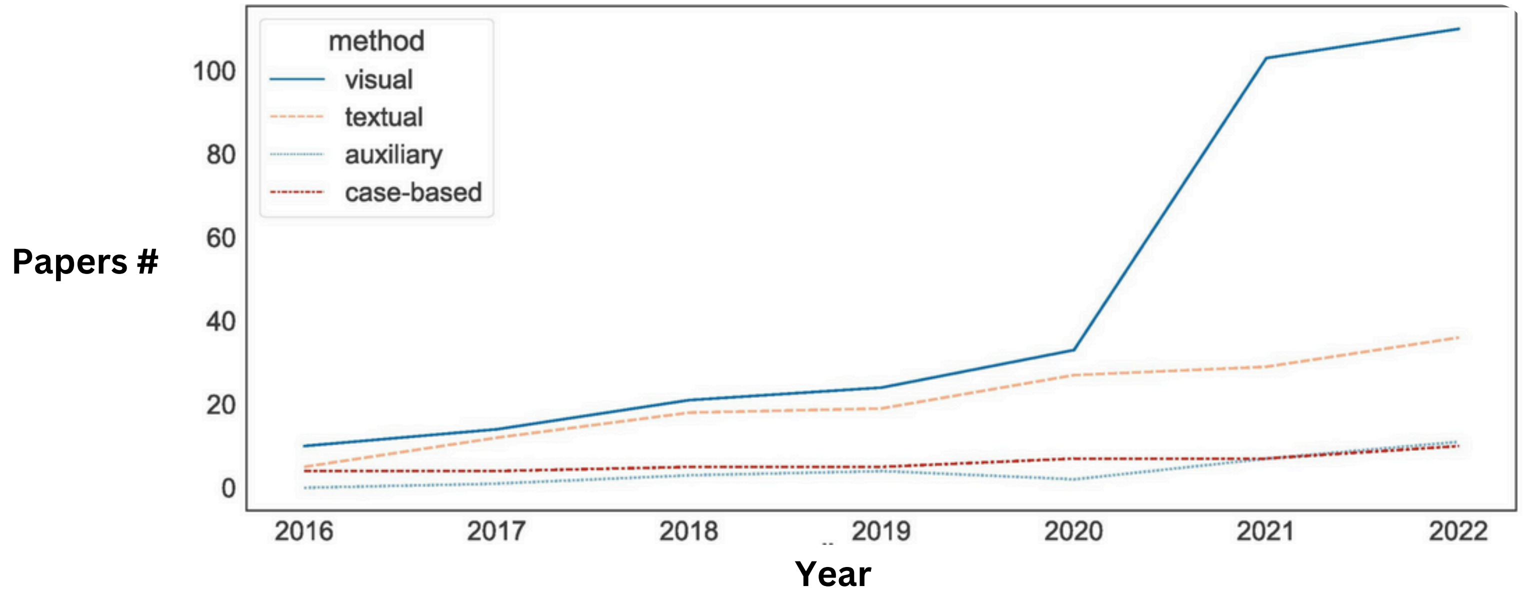


Figure 4: Number of Published Articles on XAI in Medical Image Analysis and from 2016 to 2022

- **Scope:**

1. Local: Explains individual predictions.
2. Global: Explains the overall model behavior.

- **Model:**

1. Model-Based: Models with built-in explainability, easy to interpret (e.g., decision trees).
2. Post hoc: Methods applied after model construction, treating it as a black-box (e.g., DNN).

- **Types of XAI Methods:**

1. Visual Explanations
2. Textual Explanations
3. Case-based Explanations
4. Auxiliary Explanations

• Class Activation Mapping (CAM) [1]:

$$F_k = \sum_{x,y} f_k(x,y)$$

$$S_c = \sum_k w_k^c F_k$$

$$M_c(x,y) = \sum_{k=1}^{n} w_k^c \cdot f_k(x,y)$$

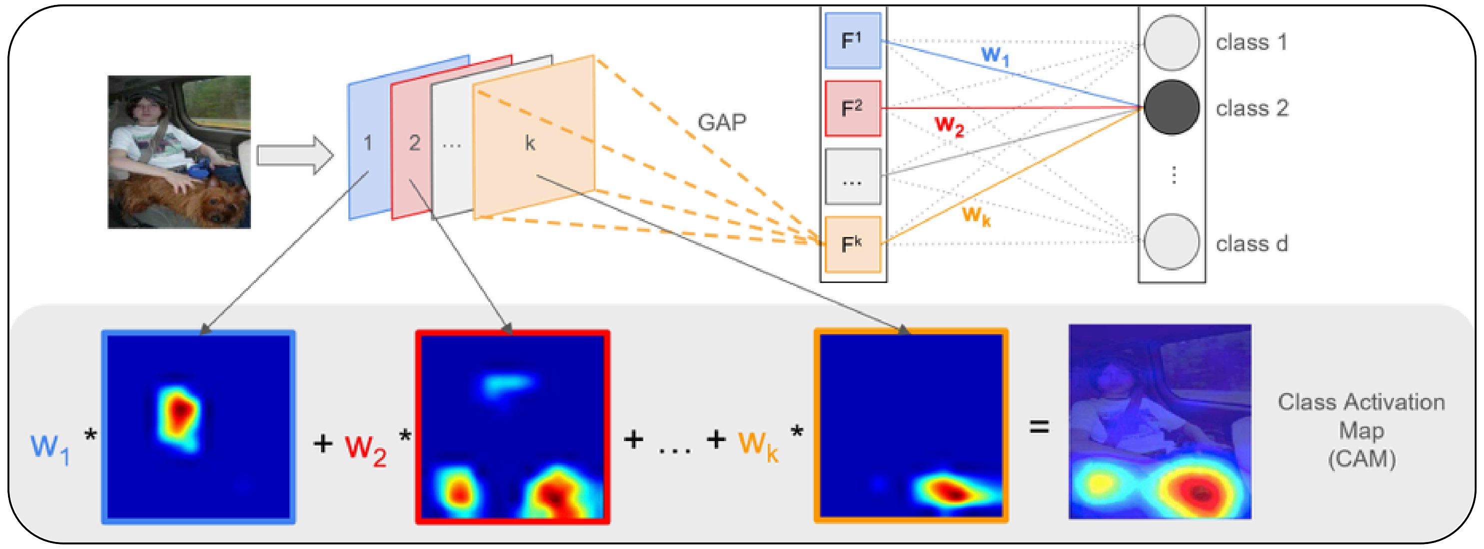


Figure 5: Class Activation Mapping (CAM) method

• Grad-CAM [2]:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y_c}{\partial A_{ij}^k}$$

$$ReLU(x) = \max(0, x)$$

$$L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right)$$

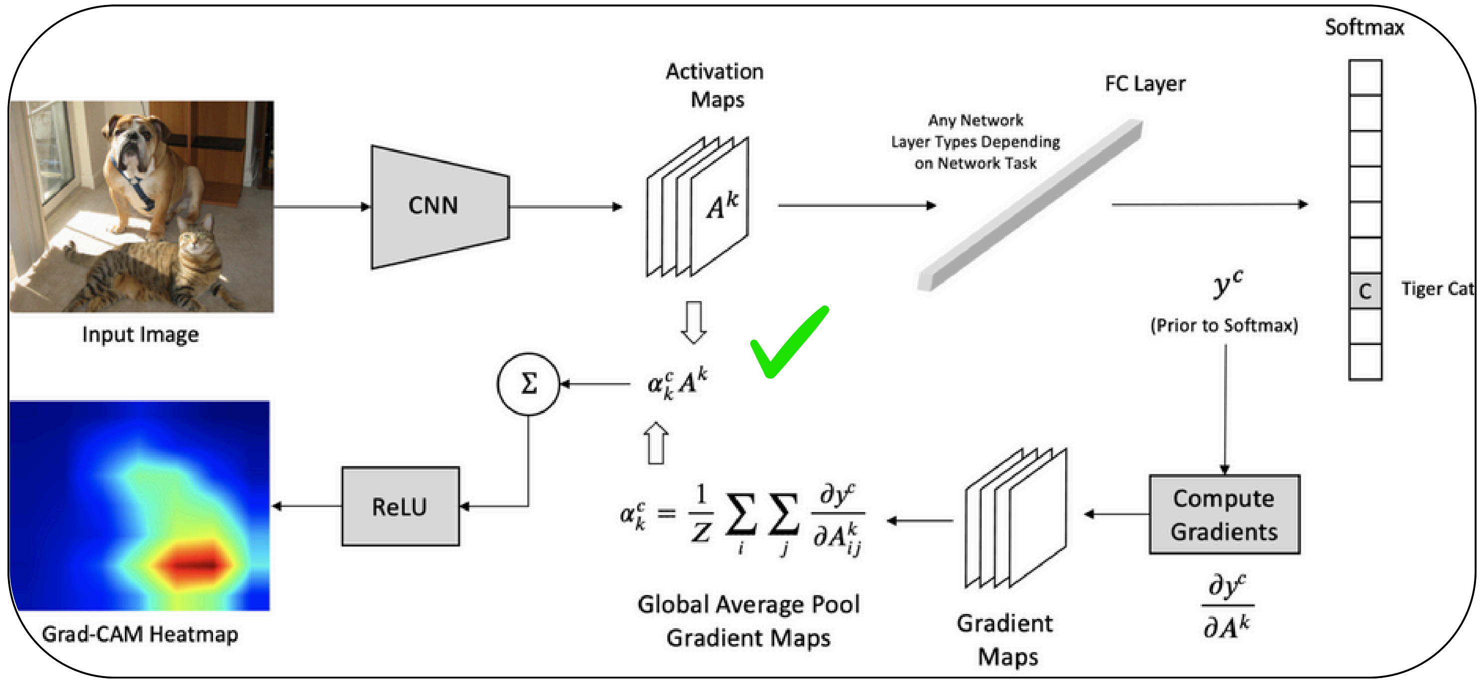


Figure 6: Gradient-weighted Class Activation Mapping (Grad-CAM) method

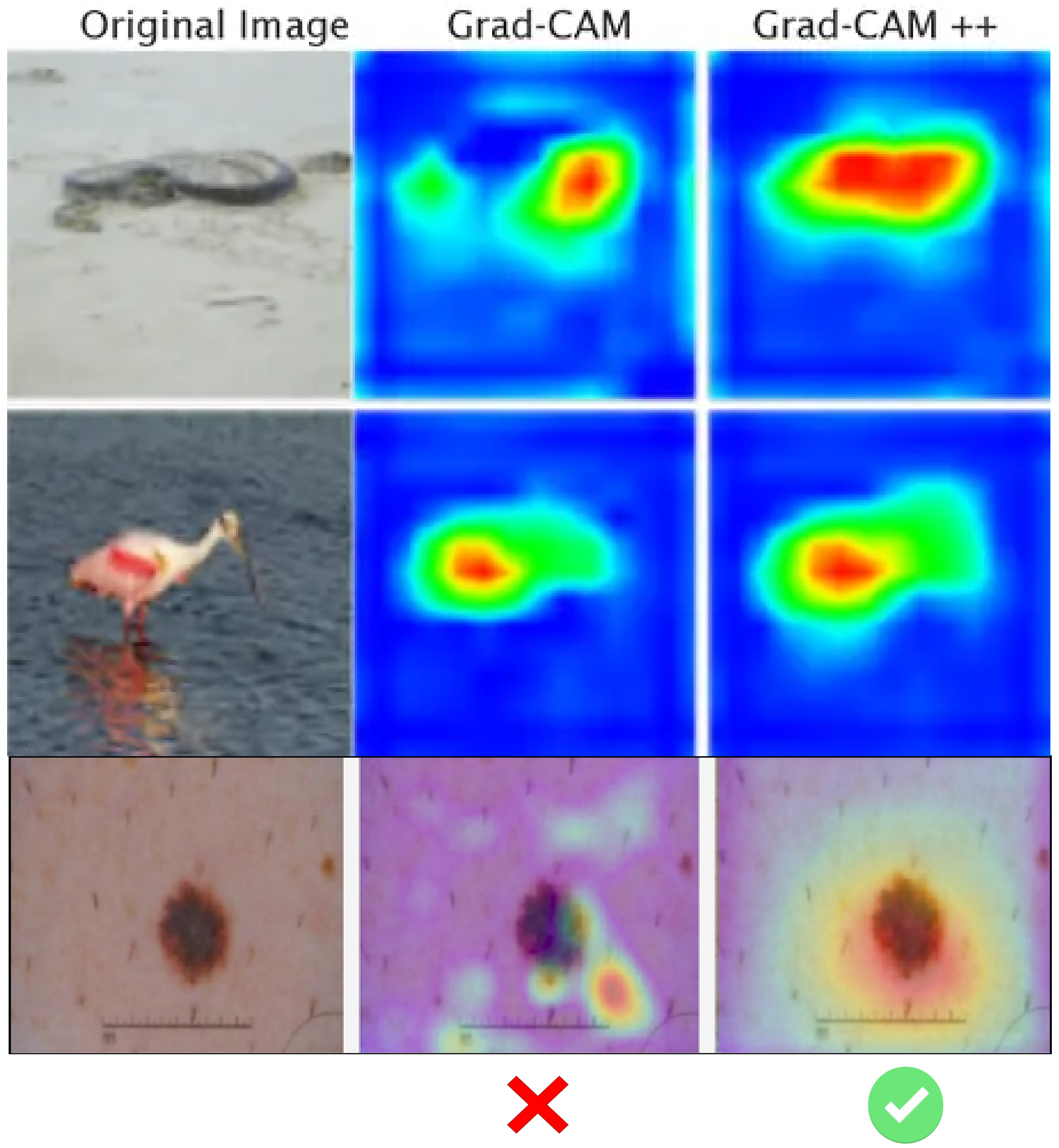
Visual Explanations

- **CAM-Based Methods:**

1. Grad-CAM
2. Grad-CAM++
3. XGrad-CAM
4. Ablation-CAM
5. FullGrad
6. Score-CAM
7. Eigen-CAM

- **Which XAI method provides better explanation by highlighting key regions in images?**

- **Better localization of objects**



XAI methods can be considered weakly supervised localization techniques in certain contexts.

- **Explainable AI (XAI) Evaluation:**

1. Human-Based
2. Computational/AI-Based Evaluation

Human-Based Evaluation Requires <u>Experts</u>, Who Are Not Always Accessible, and <u>Ground Truth</u>, Which Is Often Unavailable in Many Cases.

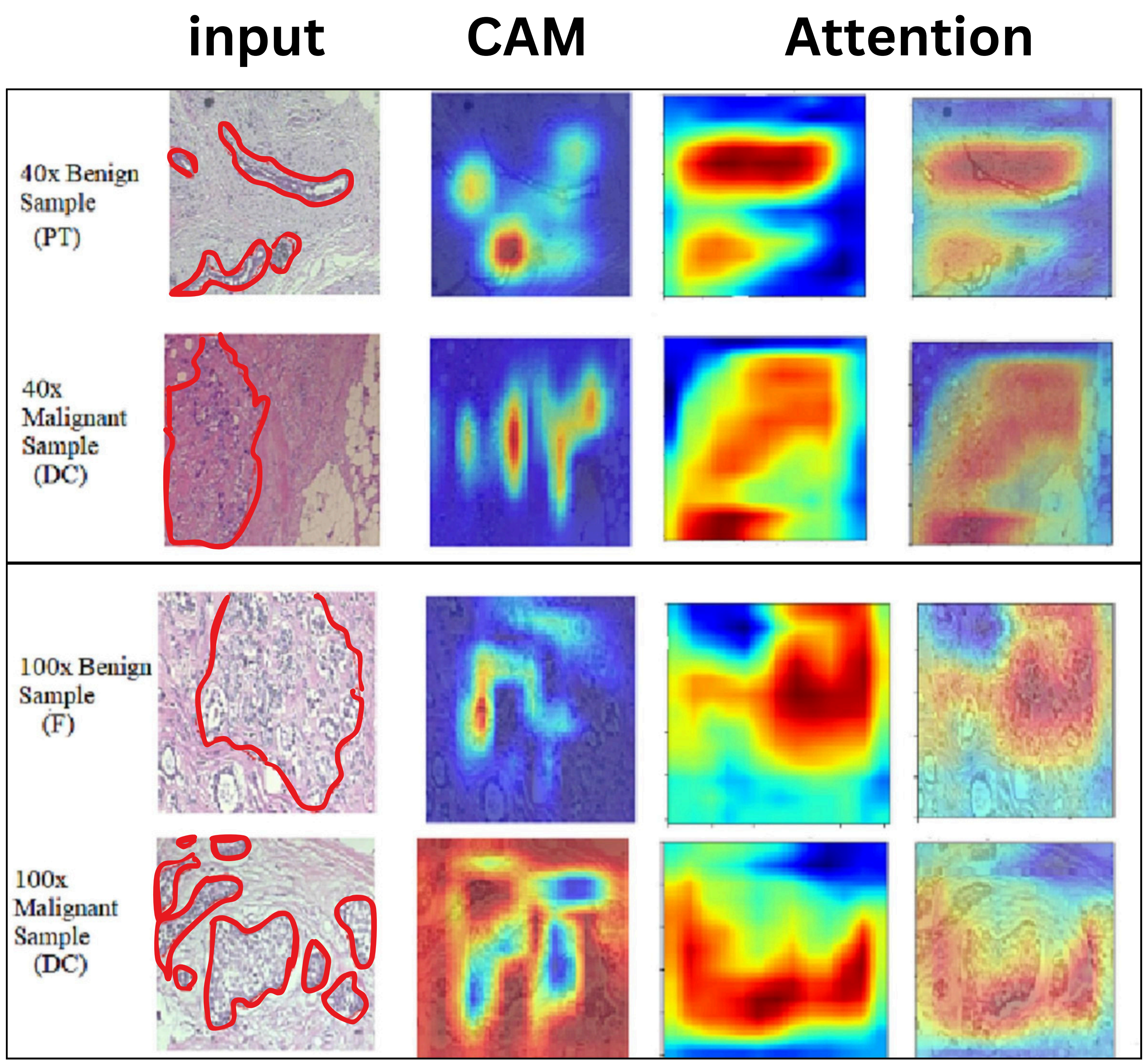


Figure 7: A Comparison of the Generated Heatmap [3]

- **Deletion Metric [4]:**

1. Most Relevant First

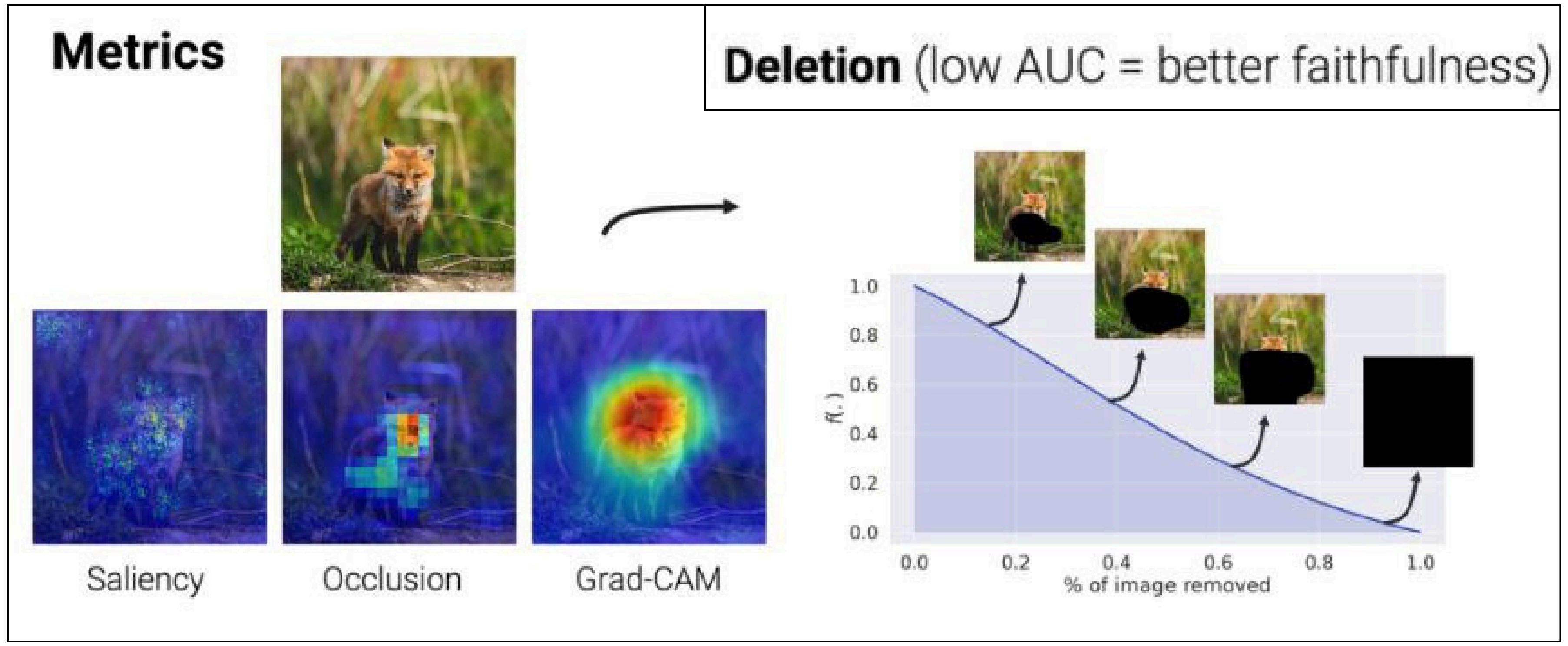


Figure 8: Deletion steps.

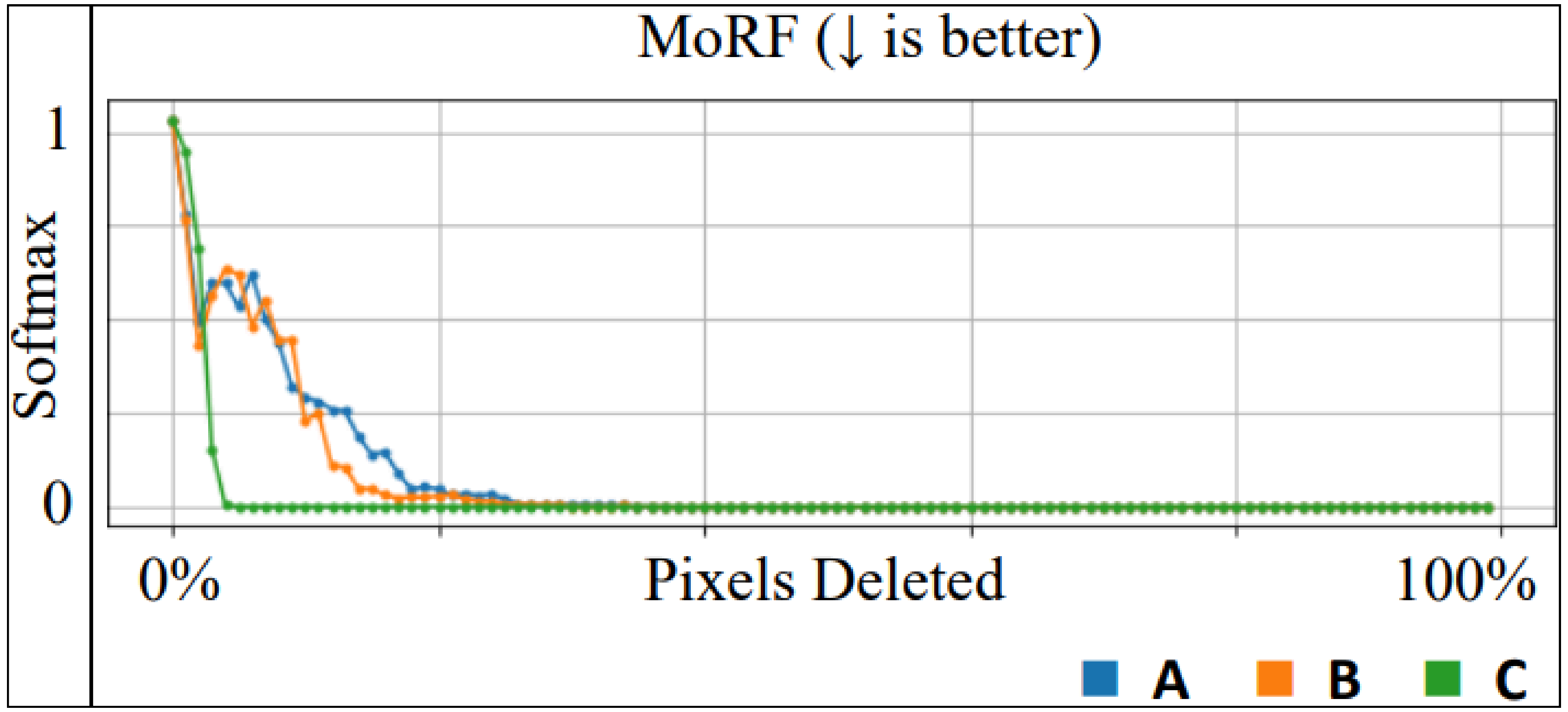


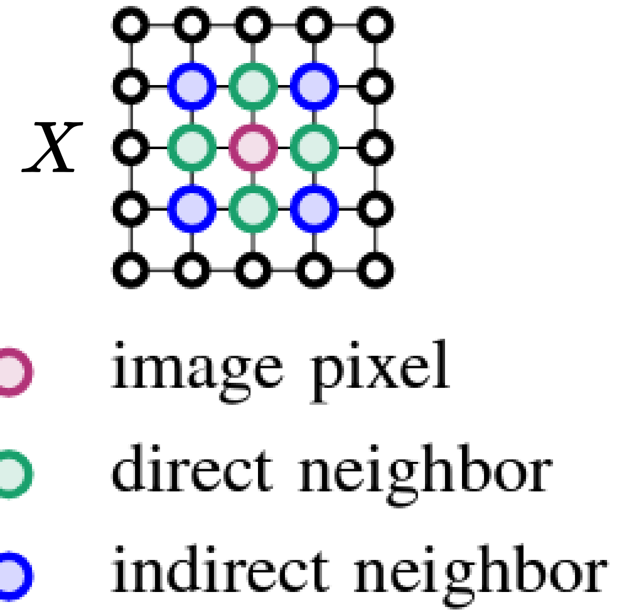
Figure 9: Comparison of the Impact of Feature Removal in Methods A, B, and C Using MoRF.

• Occlusion Strategies:

$$M = \begin{bmatrix} 1 & 0 & 0 & 1 & \dots & 1 \\ 0 & 1 & 1 & 0 & \dots & 0 \\ 1 & 1 & 0 & 1 & \dots & 1 \\ 0 & 1 & 1 & 0 & \dots & 1 \end{bmatrix}$$

1. **Blackening [5]:** $I' = I \odot (1 - M)$
2. **Blurring [6]:** $I' = I \odot (1 - M) + (I * G) \odot M$
3. **Mean [5]:** $I' = I \odot (1 - M) + \mu \odot M$
4. **Histogram [7]:** $I' = I \odot (1 - M) + H \odot M$
5. **Noisy Linear Imputation [8]:**

$$x' = (1 - M) \odot x + M \odot (X\hat{\beta} + \epsilon), \hat{\beta} = \frac{1}{6}, \frac{1}{12}$$



Challenges

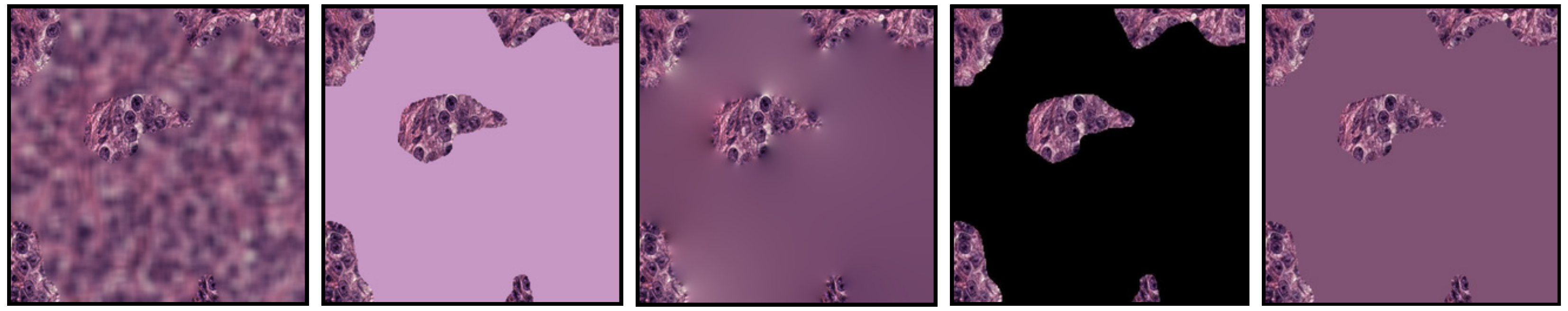
Challenges

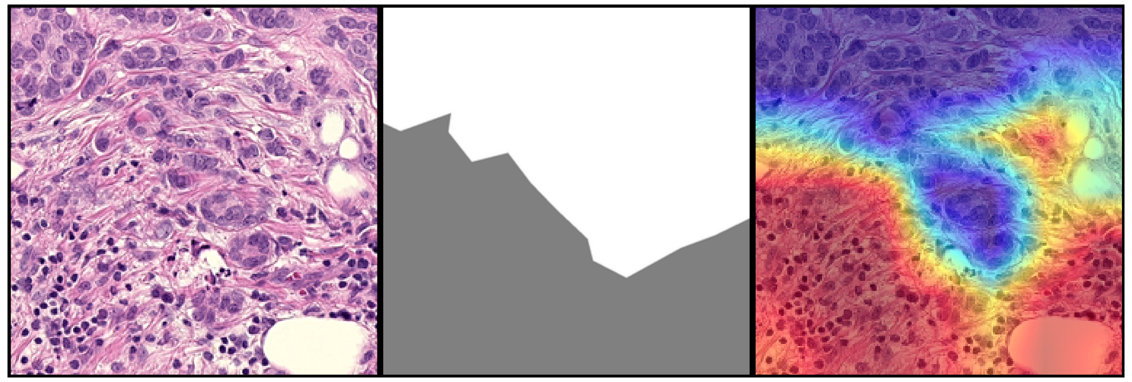
• Challenges of Occlusion Strategies:

- **Key Issues with Artificial Modifications:**

1. **Unrealistic Changes:** May generate OoD samples or artifacts on images.
2. **Incorrect Evaluation:** Removing unimportant features can mislead model performance.
3. **Medical Sensitivity:** Hinders detection of critical features, impacting diagnosis.
4. **Risk of Misdiagnosis:** Errors in feature evaluation may lead to diagnostic mistakes.

"Evaluation of XAI methods through Deletion metric is highly dependent on how the features are removed."





Tumor patch | Binary mask | EigenCAM Heatmap

Tumor probabilities - EigenCAM - Blackening :

[0.99989, 0.438763]

Tumor probabilities - EigenCAM - Blur :

[0.99989, 0.90]

Trade-off between deleting the feature and preserving the distribution.

Sharif University of Technology

Proposed Methodology

01

02

03

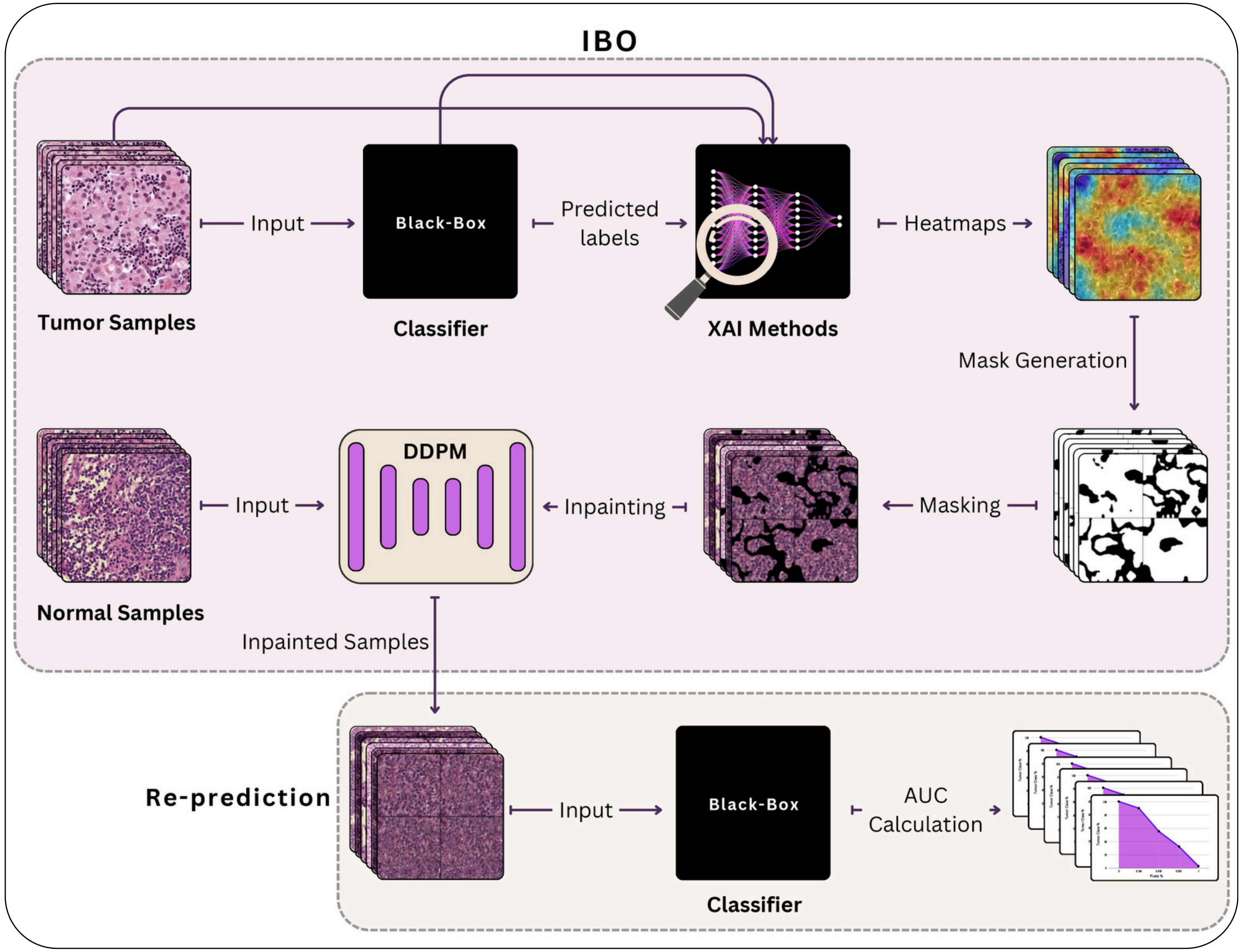


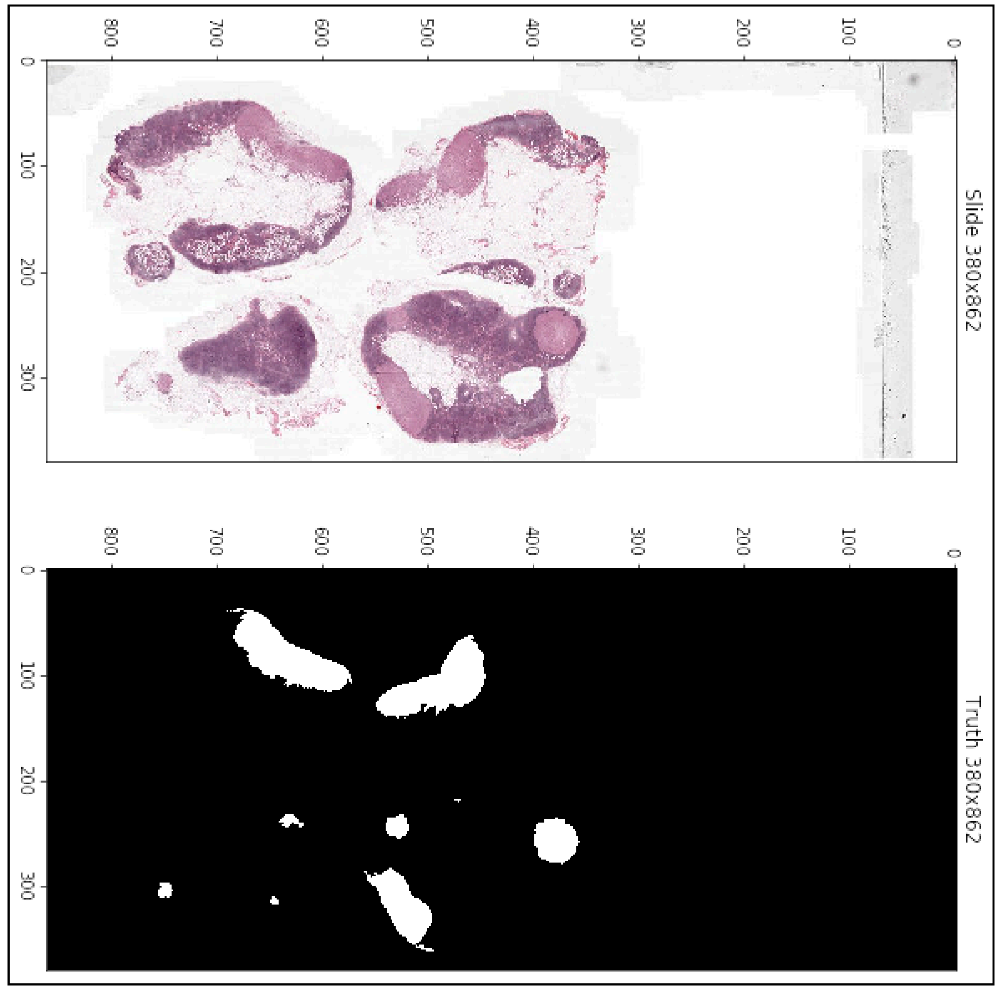
Figure 10: Proposed Methodology

• Steps:

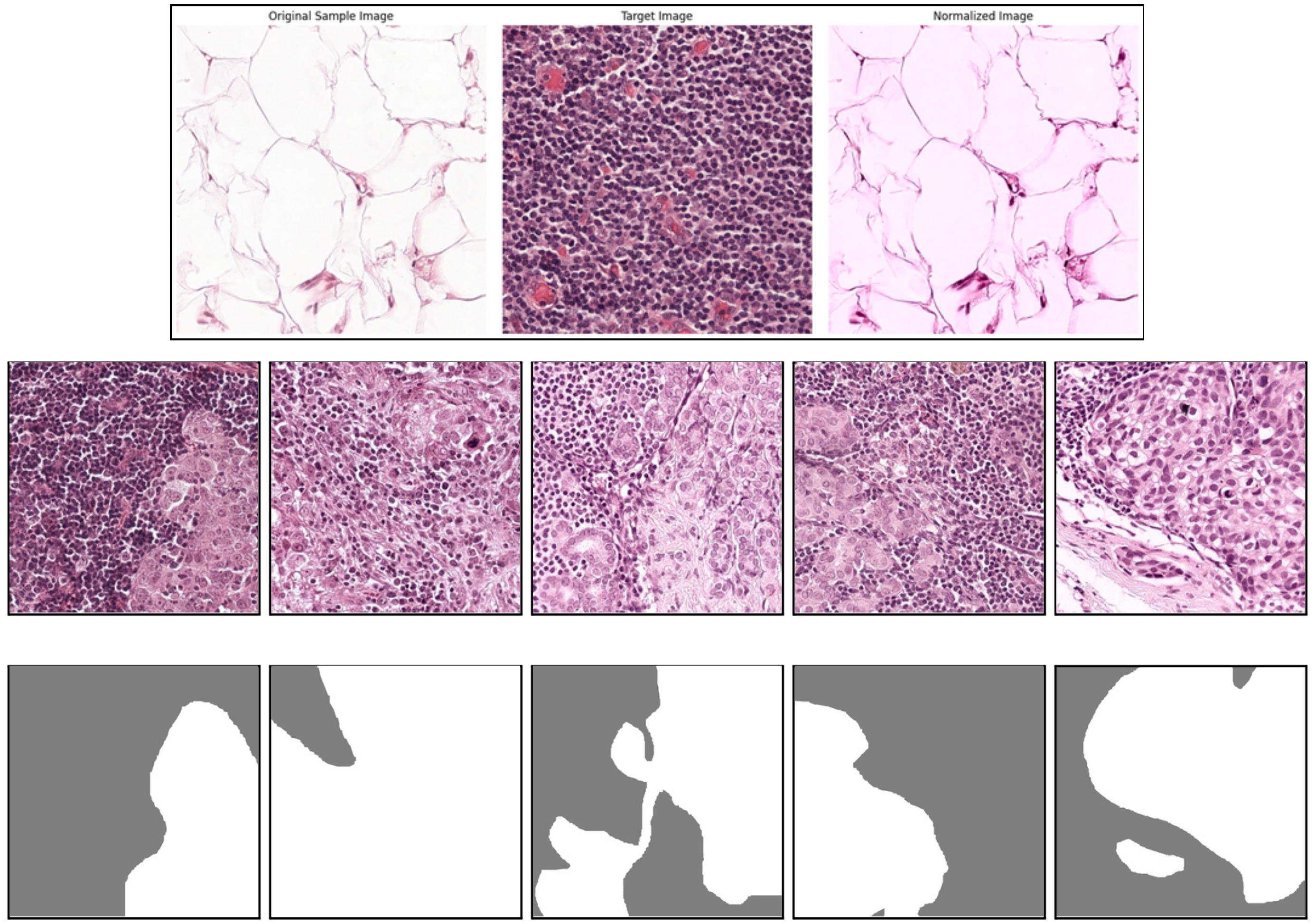
1. Training a Classifier
2. Generating Heatmaps (Using CAM-Based Methods in This Study)
3. Generating Masks Based on Heatmaps
4. Image Inpainting
5. Re-prediction
6. AUC Calculation and Analysis

CAMELYON16 [10]

- 400 WSI image of Sentinel Lymph Nodes. (2 Classes: Tumor/Normal)
- 270 images with precise annotations by pathologists and specialists.



"Normal Staining and 512x512 Patching Applied"



- In this step, a total of <u>15,214</u> samples were extracted for each of the tumor and healthy/normal classes.

- **Training VGG16 and ResNet50 for 20 Epochs:** ResNet50 Accuracy **98.62%**, VGG16 Accuracy **96.43%**.

ResNet50 Was Selected for Further Work.

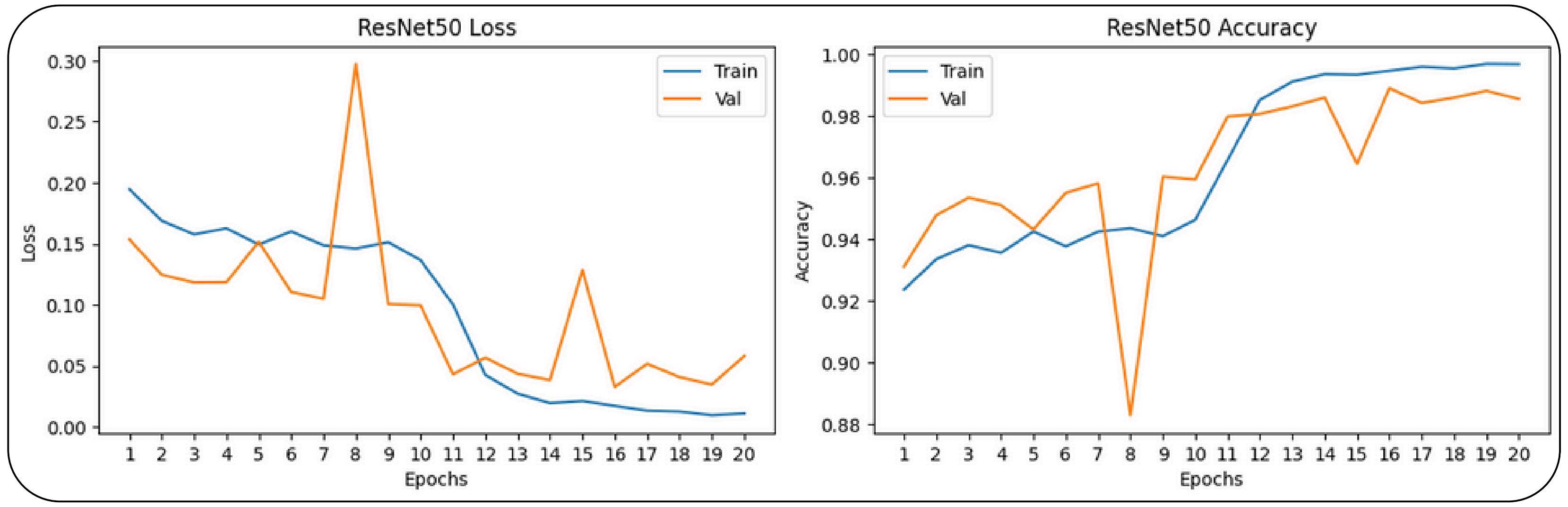
- batch_size_train = 32
- optimizer = Adam
- Learning rate = 0.0001
- Total dataset length: 30428 (.7, .15, .15)
- Train dataset length: 21299
- Validation dataset length: 4564
- Test dataset length: 4565

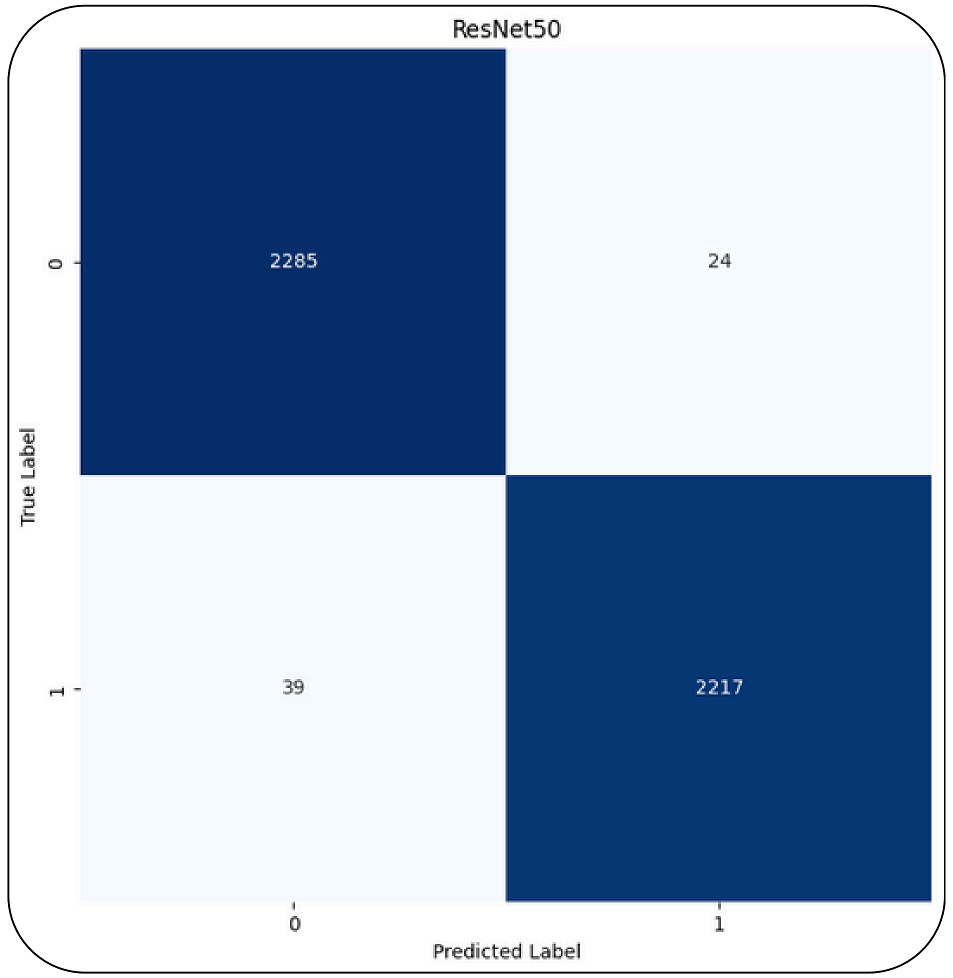
- Kaggle Platfrom - T4 x 2

```
Classification Report:
              precision    recall  f1-score

           0     0.9832    0.9896    0.9864
           1     0.9893    0.9827    0.9860

    accuracy                         0.9862
```

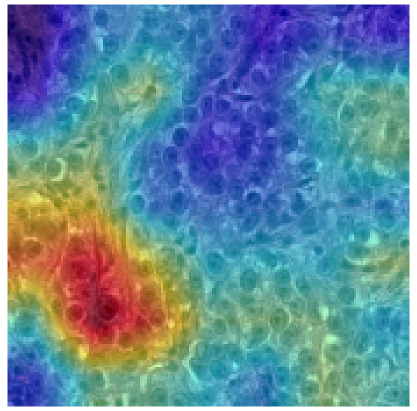




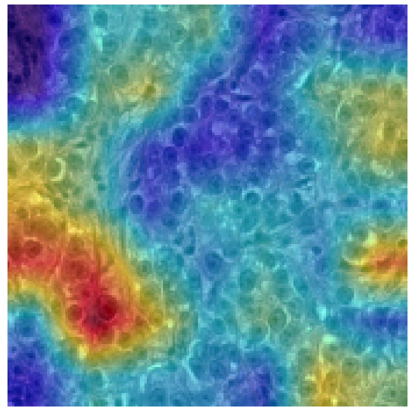
Generating Heatmaps

- **7 CAM-Based Methods Were Selected and Applied to Tumor Patches.**

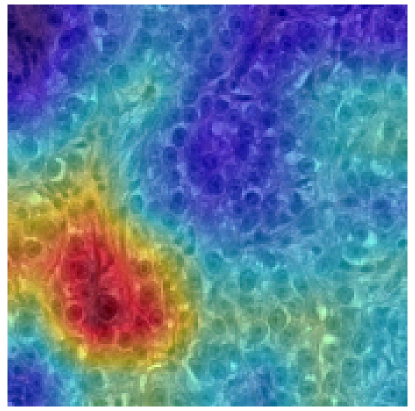
1. Grad-CAM
2. Grad-CAM++
3. XGrad-CAM
4. Ablation-CAM
5. FullGrad
6. Score-CAM
7. Eigen-CAM



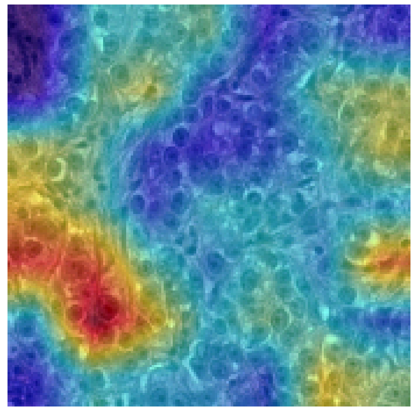
Ablation-CAM



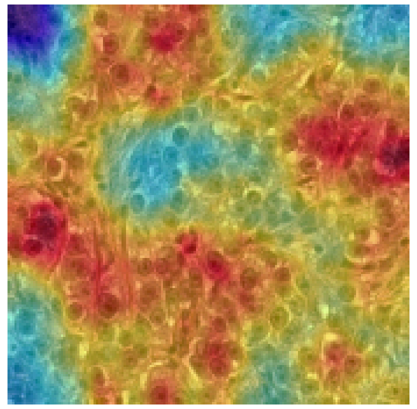
XGrad-CAM



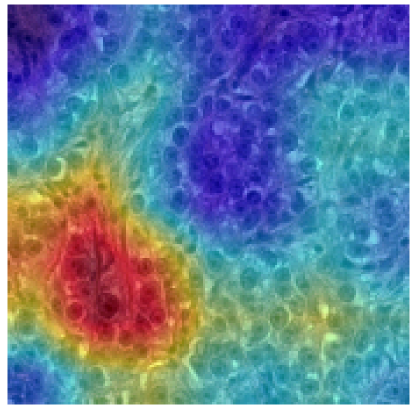
Grad-CAM++



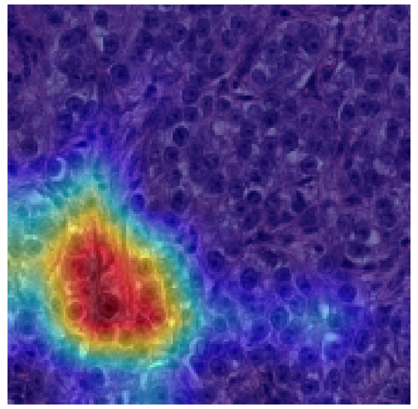
Grad-CAM



Full-Grad



Score-CAM



Eigen-CAM

Generating Masks Based on Heatmaps

- **K-means Clustering Based on Pixel Intensity Was Used for Mask Generation.**

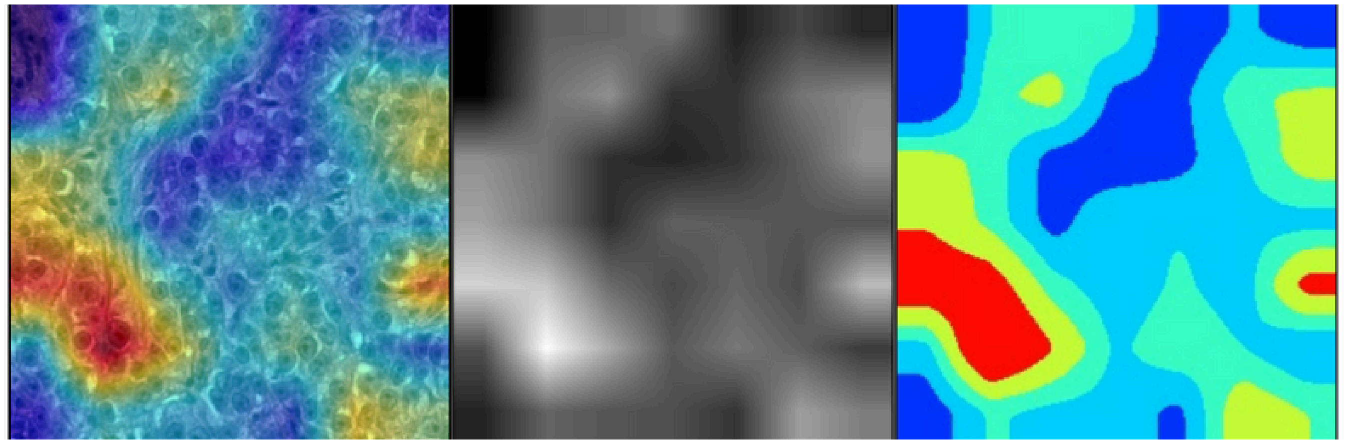
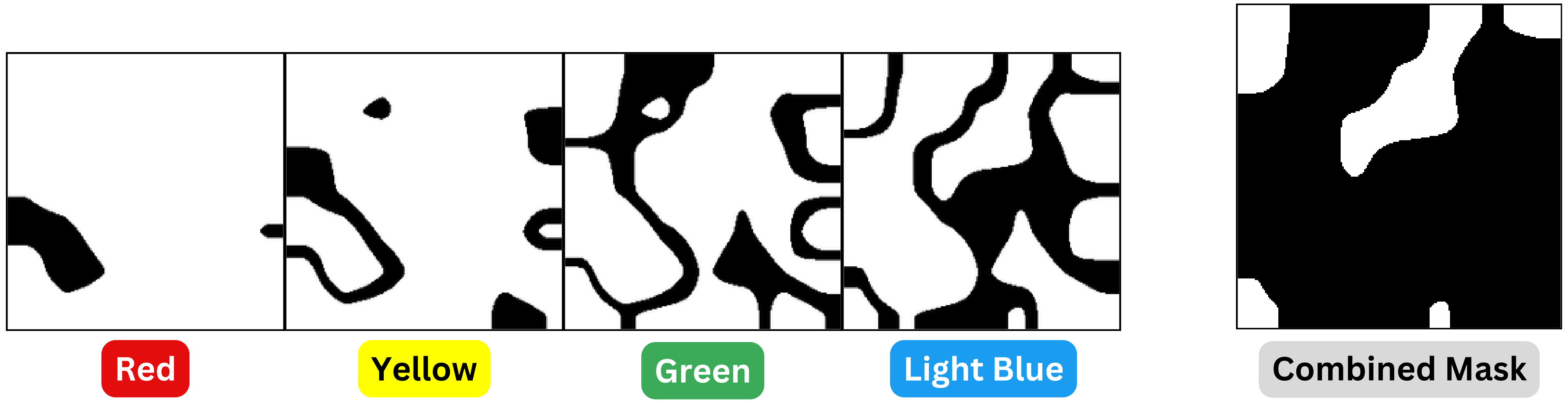


Figure 11: Comparison of Generated Heatmap, Grayscale Heatmap, and Importance Levels

Red Areas: Most important for model decision-making
Yellow Areas: Highly important
Green Areas: Moderately important
Light Blue Areas: Less important
Blue Areas: Least important, removed during inpainting process

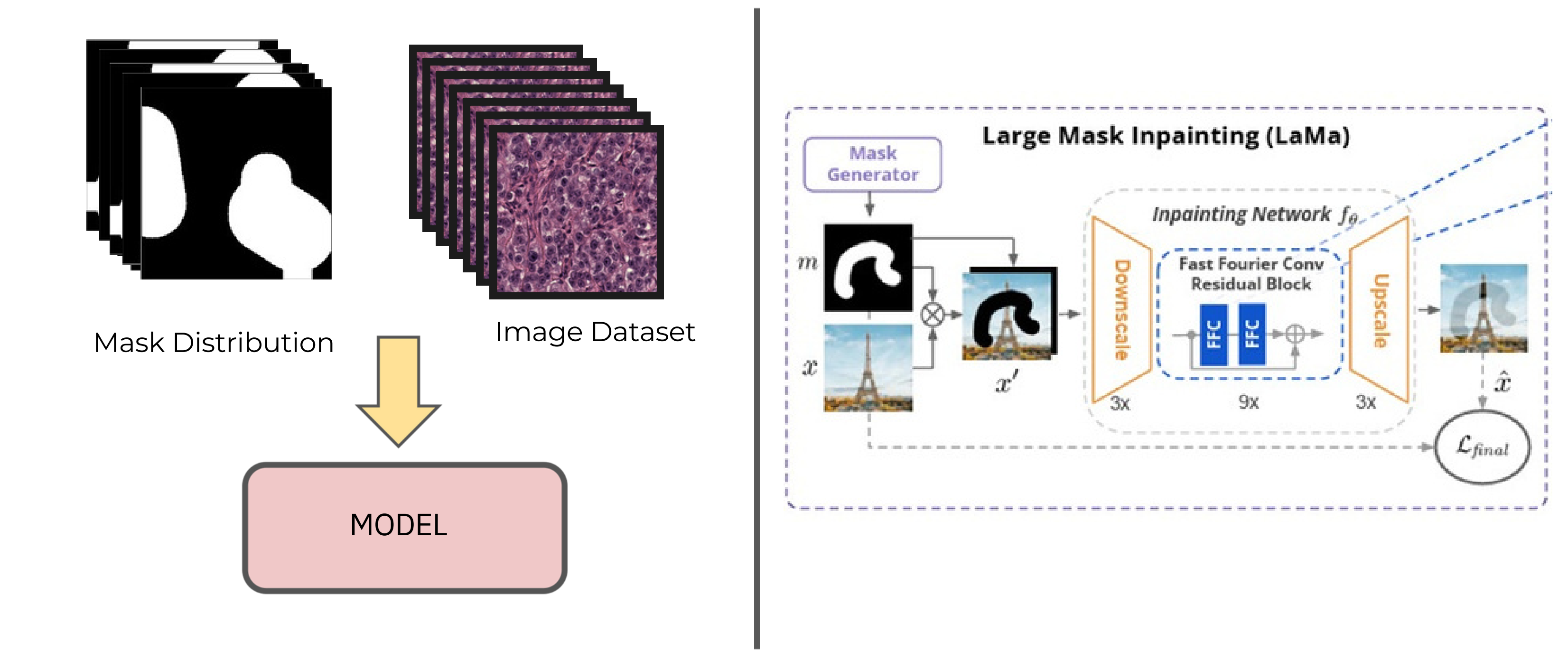
$$J = \sum_{i=1}^{k} \sum_{x_j \in C_i} \|x_j - \mu_i\|^2, k = 5$$



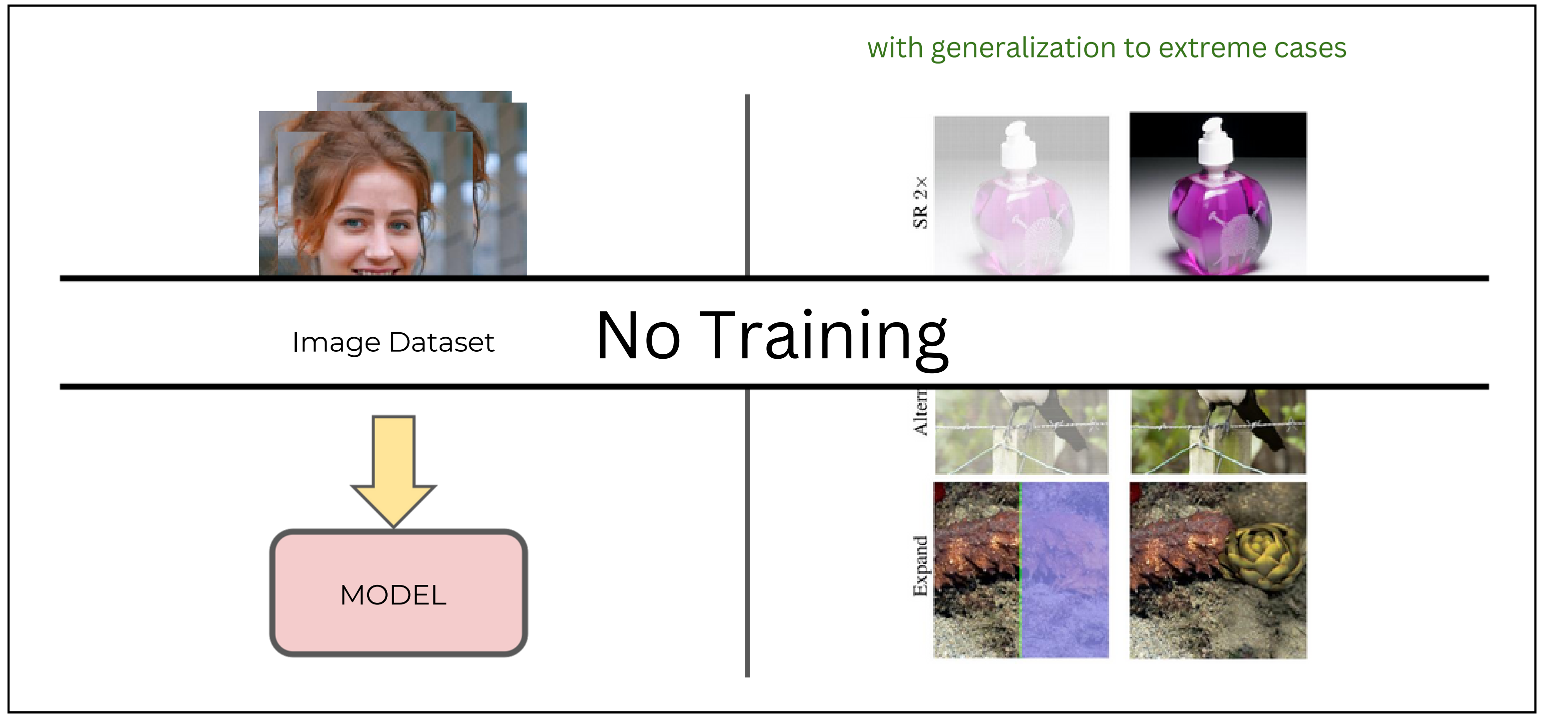
- **The Black Areas indicate the regions that need to be inpainted.**
- **4 Inpainting Steps.**

Inpainting

- **Existing approaches train with a given mask distribution**



- **Repaint (2022) [11]**



- **Preliminary: DDPM**

Forward process (rewritten using independence property of noise added at each step)

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}) \longrightarrow q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I})$$

Reverse process — Neural network predicts

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \mathrm{Uniform}(\{1, \dots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$
6: **until** converged

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \dots, 1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right) + \sigma_t\mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \ldots, 1$ **do**
3: $\quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\quad \mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right) + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

$x_T \sim \mathcal{N}(0, I)$
for $t = T$ **to** 1 **do**
$\quad$ **for** $u = 1$ **to** U **do**
$\quad\quad \epsilon \sim \mathcal{N}(0, I)$ **if** $t > 1$, **else** $\epsilon = 0$
$\quad\quad x_{t-1}^{\text{known}} = \sqrt{\bar{\alpha}_t}x_0 + (1 - \bar{\alpha}_t)\epsilon$
$\quad\quad z \sim \mathcal{N}(0, I)$ **if** $t > 1$, **else** $z = 0$
$\quad\quad x_{t-1}^{\text{unknown}} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right) + \sigma_t z$
$\quad\quad x_{t-1} = m \odot x_{t-1}^{\text{known}} + (1 - m) \odot x_{t-1}^{\text{unknown}}$

Known and Unknown

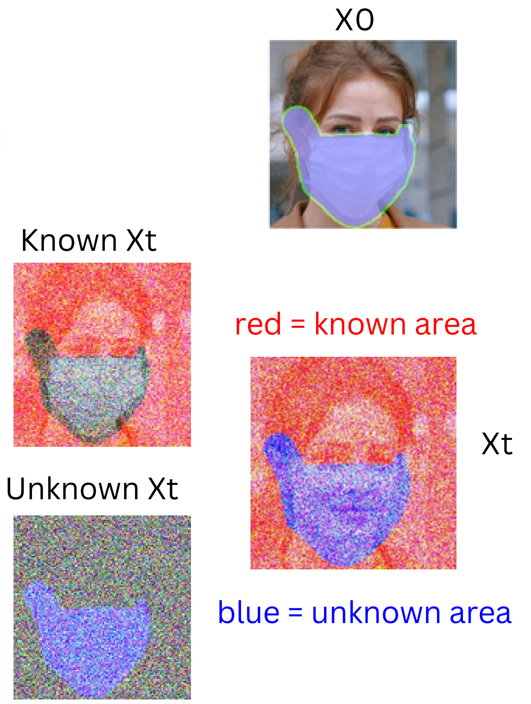
$$x_{t-1} = m \odot x_{t-1}^{\text{known}} + (1 - m) \odot x_{t-1}^{\text{unknown}}$$

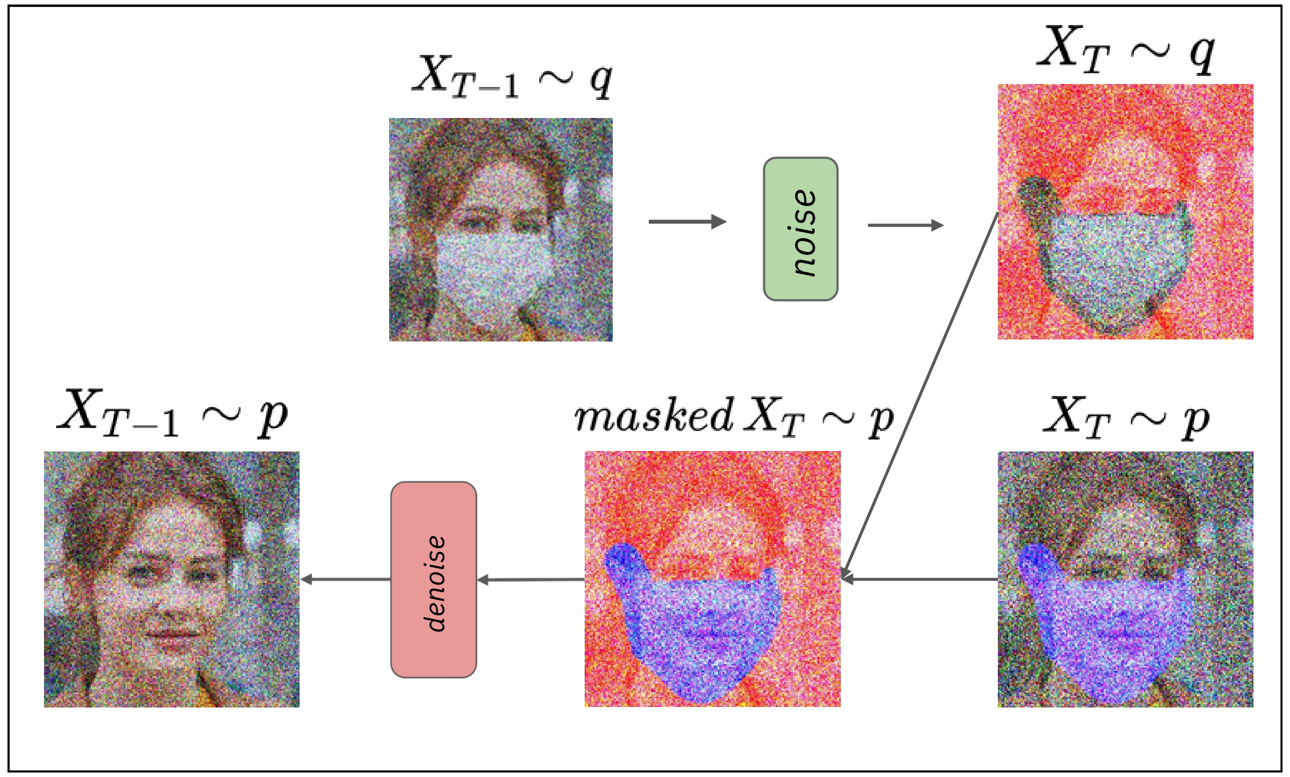
Known is obtained from Forward Process

$$x_{t-1}^{\text{known}} \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)\mathbf{I})$$

Unknown is obtained from the denoise process

$$x_{t-1}^{\text{unknown}} \sim \mathcal{N}(\mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

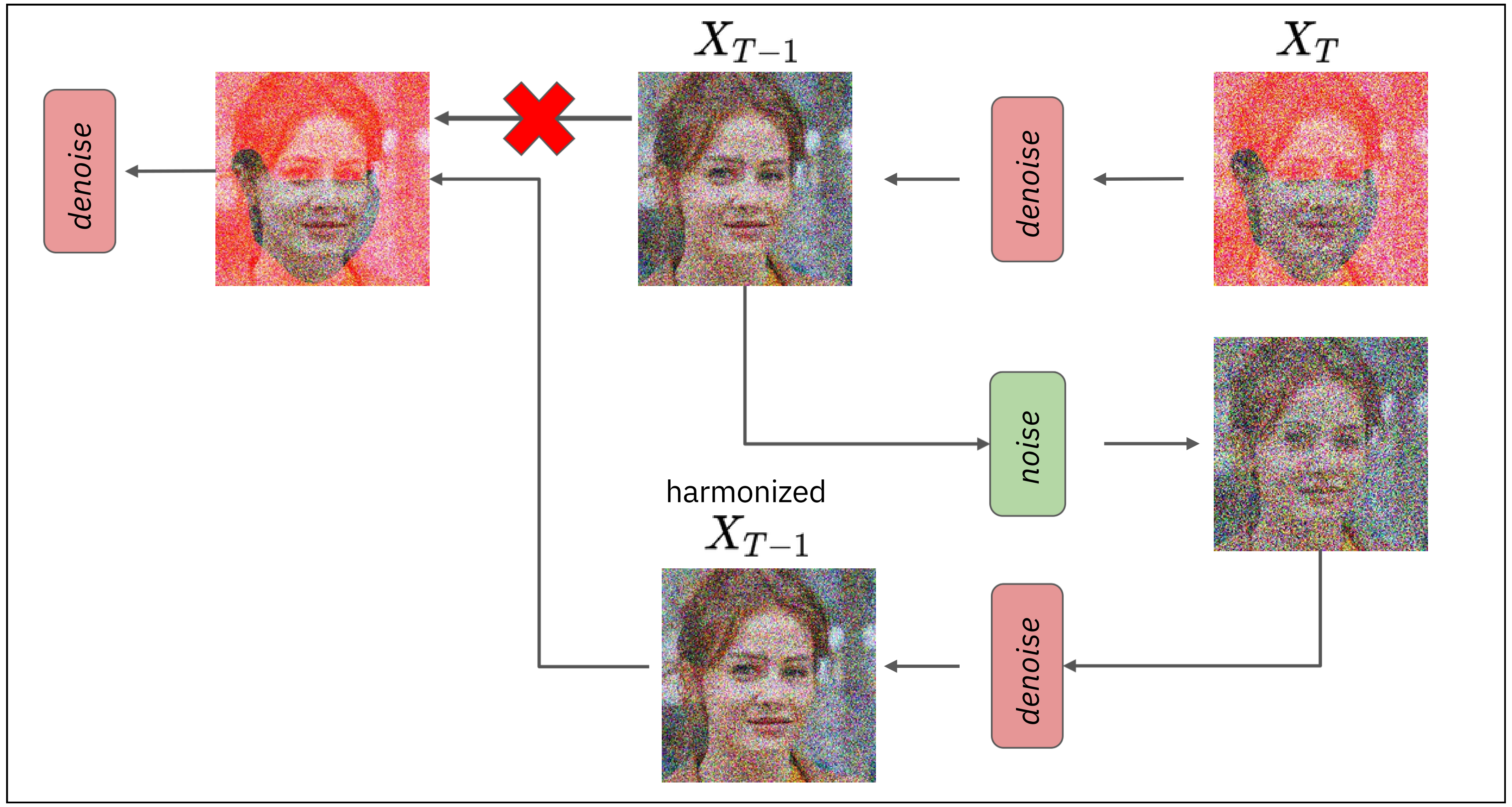




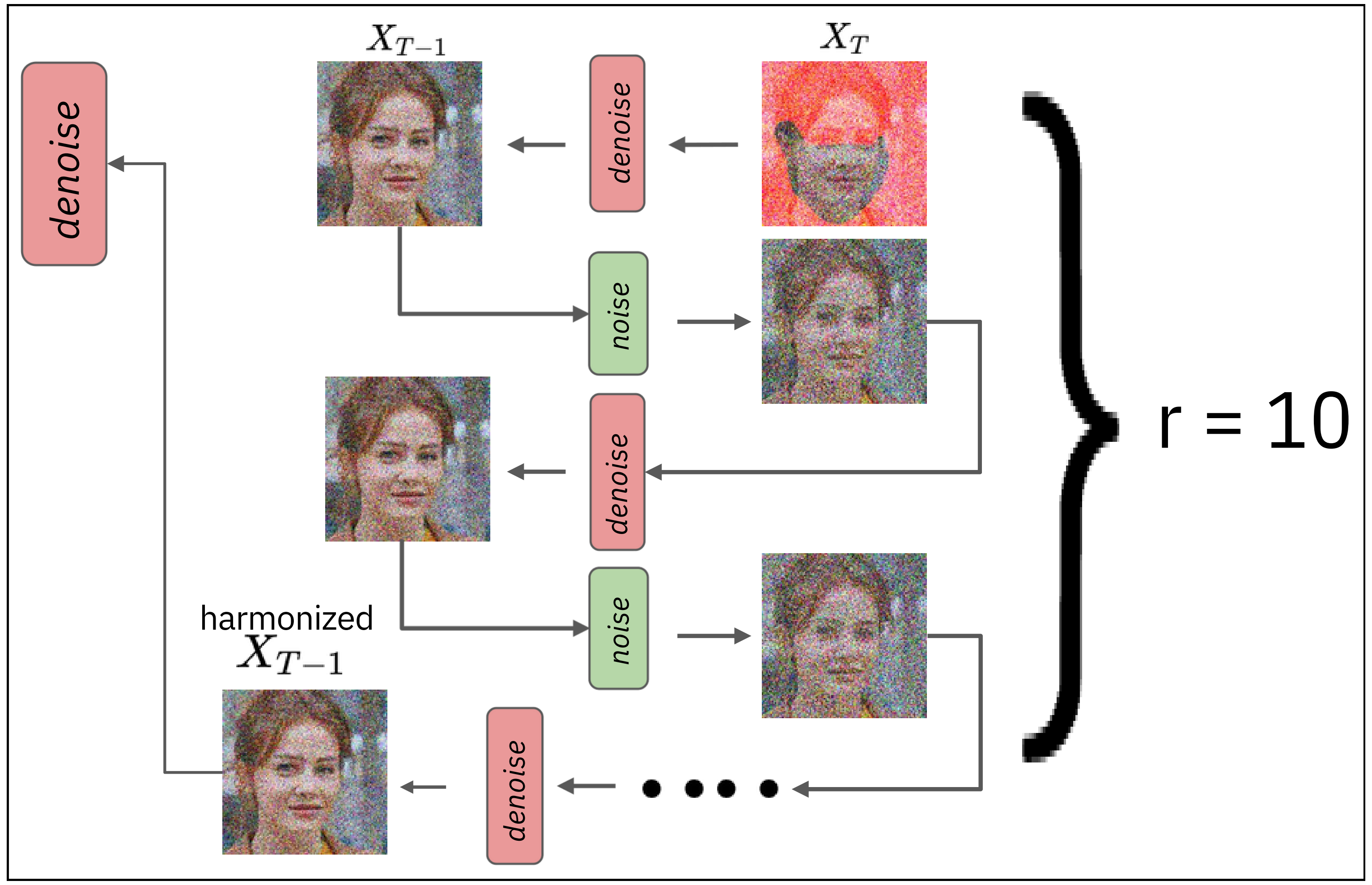
The main issue is when sample the original noise (red), it has no information about the generated part (non-red).

- Solution: Repeat the noising and de-noising steps during inference

• Resampling



Resampling (r = 1)



Resampling with more steps (r = 10)

• Jumping

- Reason:
  - Resampling at every step would make the image blur.
- Only do resampling every j time.
  - For example, when j = 10 and T = 250, only do resampling when t = 240, 230, 220, 210..., and the length of resampling would be 10.

Resampling with jump length (j = 10)

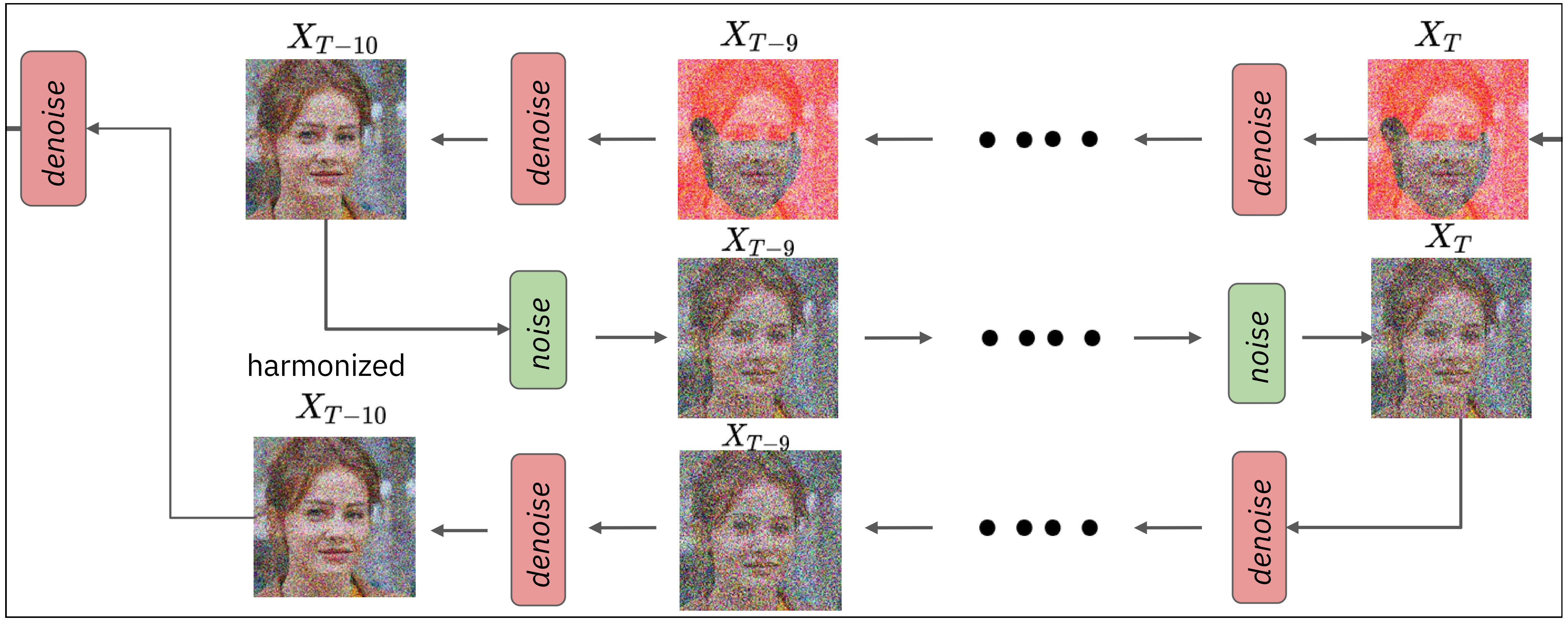
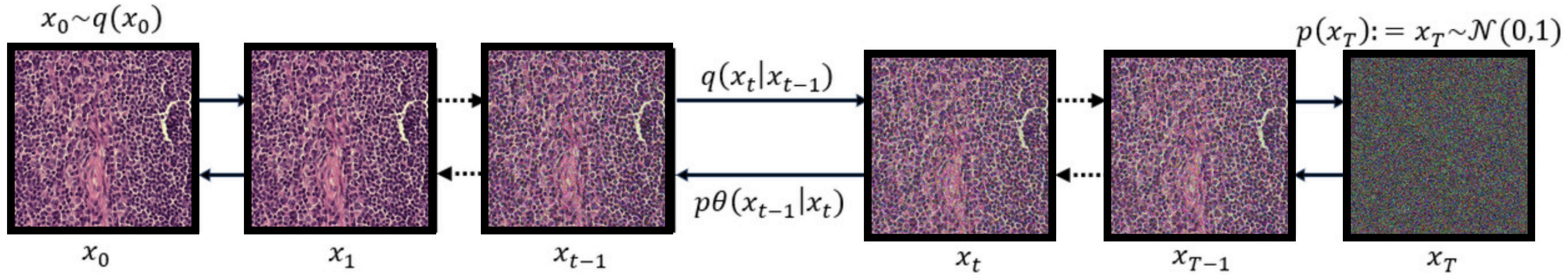
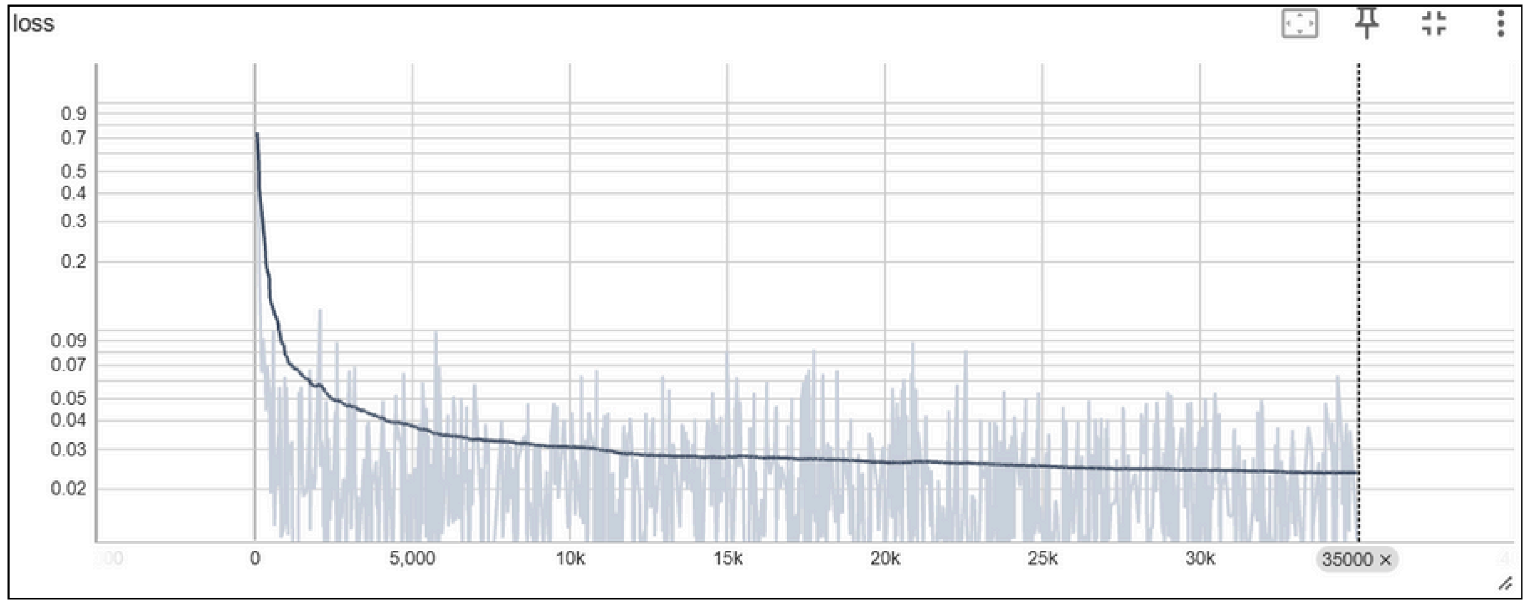


Image Inpainting

• Training DDPM on Normal Samples:

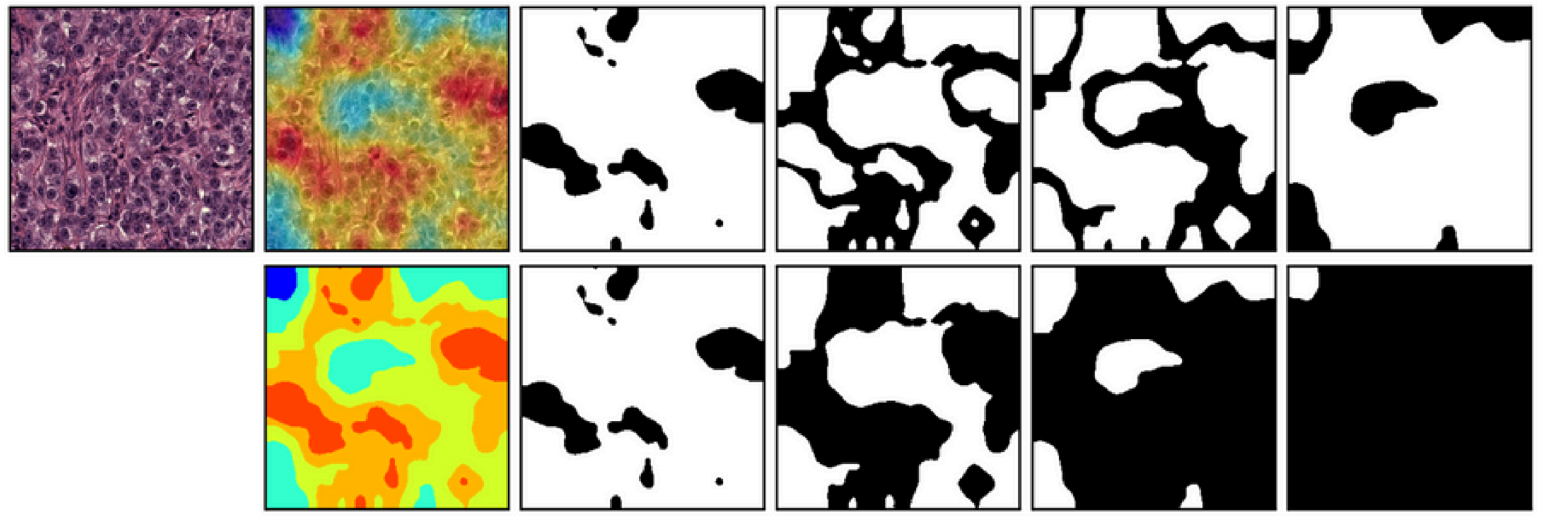


- 35000 steps.
- A100, (Colab Pro+)
- Image Size = 224 x 224
- layers_per_block = 2
- batch size = 16
- optimizer = AdamW
- Learning rate = 0.0001
- Number channels = 128, 256, 512
- Linear variance scheduler
  β in the range [0.0001, 0.02] and set the total timesteps T = 1000



DDPM loss.

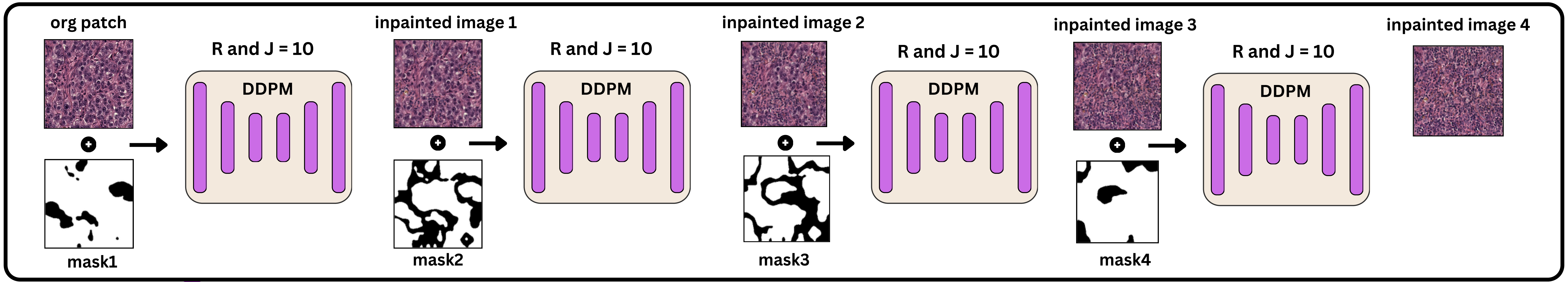
• Inpainting Steps: In 4 Steps

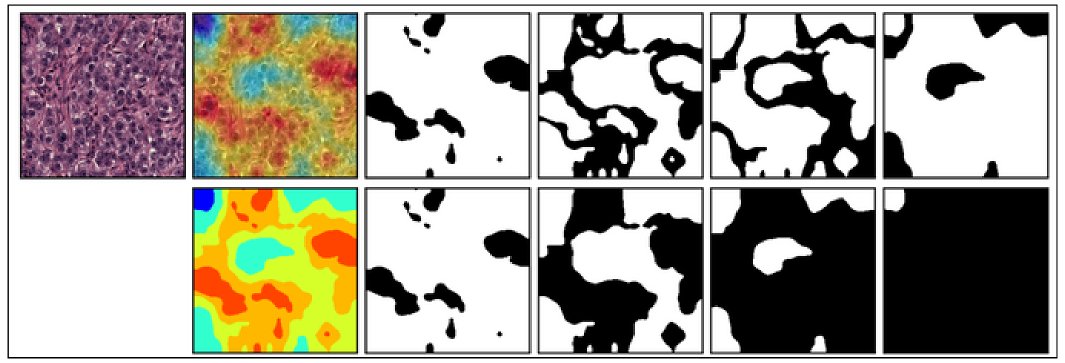


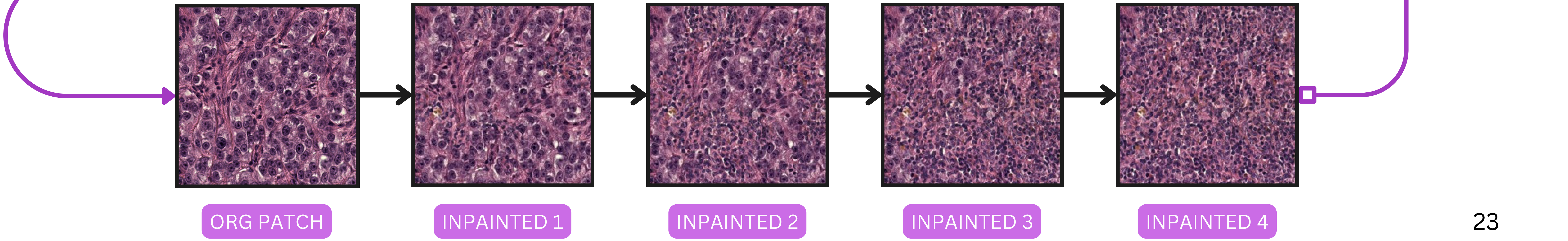
The tumor patch, its Full-Grad heatmap, and corresponding masks are shown.

The first row highlights regions with varying importance, while the second row illustrates progressive occlusion based on these regions' significance.

Inpainting Approach







AUC Calculation

Black-Box

Classifier

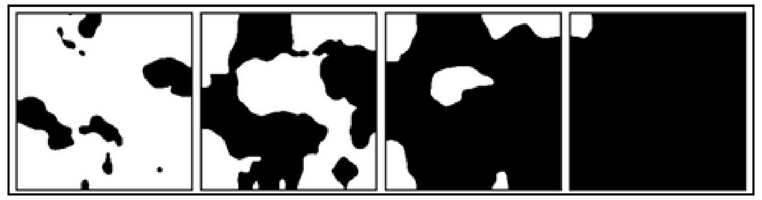
Re-Prediction

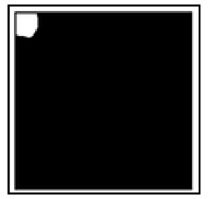
AUC Calculation

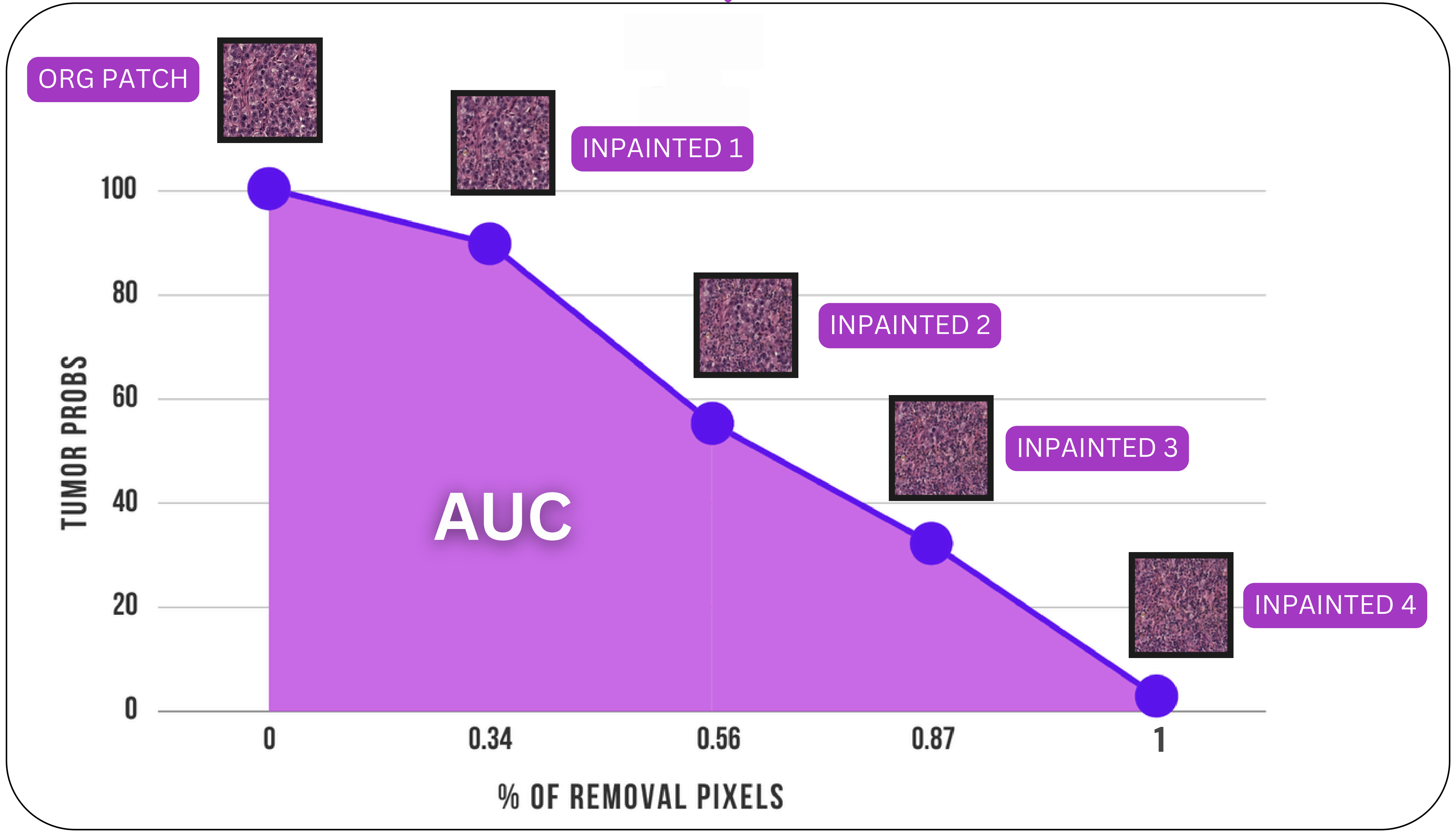
$$\text{AUC} = \int_0^1 f(p)\, dp$$

$$\text{AUC} = \sum_{i=1}^{M-1} \frac{f(p_i) + f(p_{i+1})}{2} \cdot (p_{i+1} - p_i)$$

$$p_i = \frac{\text{Number of important pixels removed in step } i}{\text{Total number of important pixels}}$$







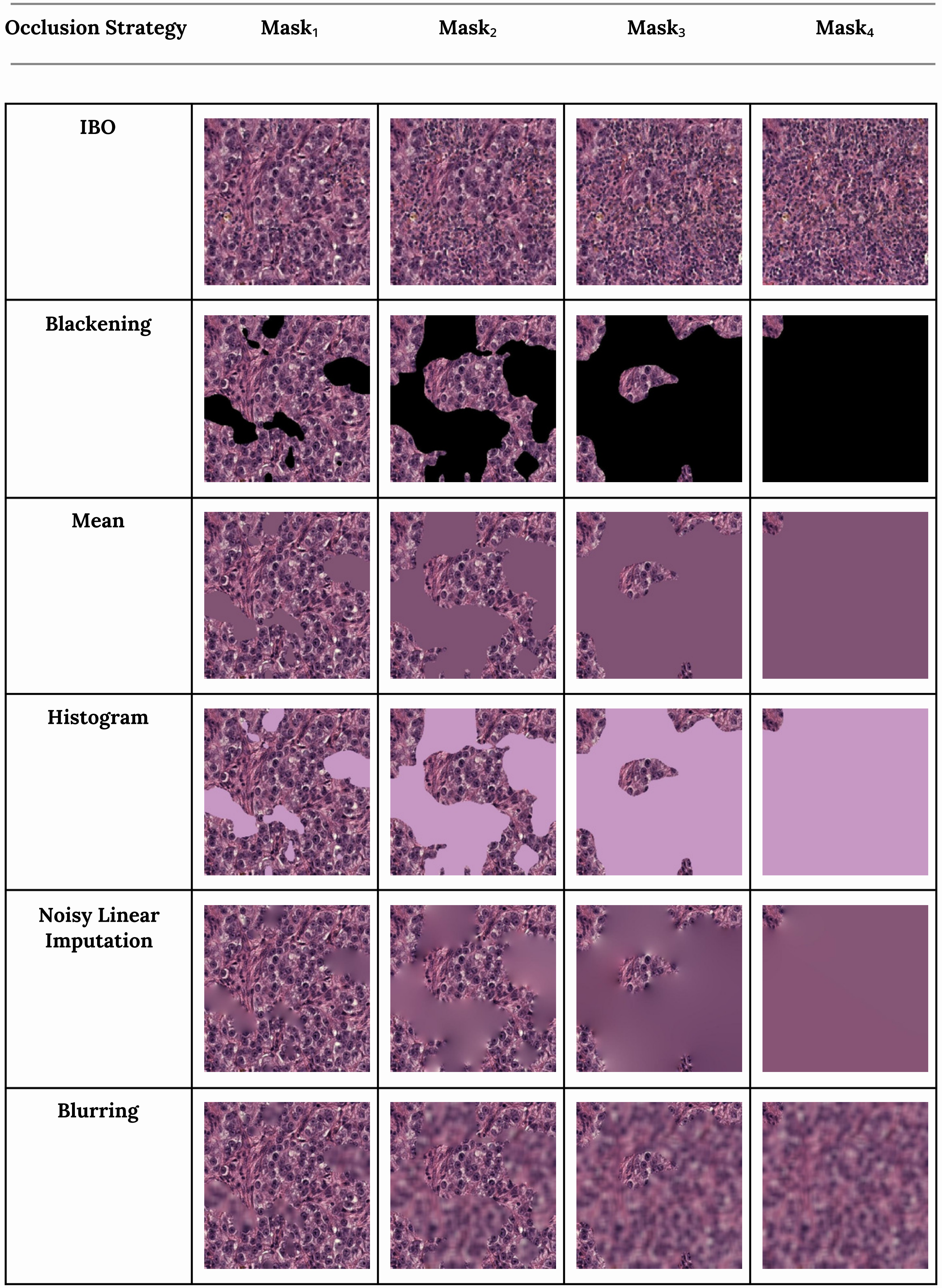
Experiences

1. For evaluating our framework and other methods, 100 random tumor patches were selected from the test set.
2. Seven XAI methods were applied to the patches.
3. Corresponding masks were generated for all patches and all XAI methods.
4. Previous occlusion methods were applied to the patches for each XAI method.
5. Our proposed method was also applied. (Inpainting Based Occlusion (IBO))

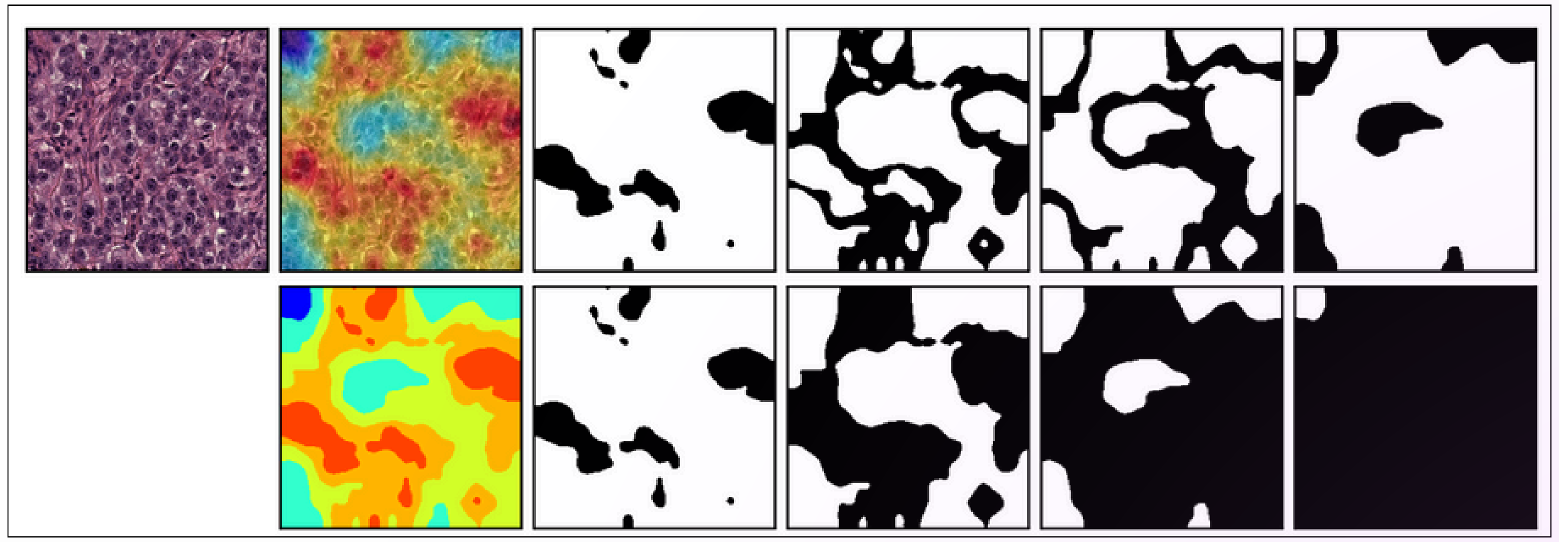
3 × A100 GPUs (Colab Pro+) were used, with each inpainting taking between 1 to 1.5 minutes.

01 Evaluation of Inpainted/Occluded Samples

02 Quantitative Evaluation



- Illustration of various occlusion strategies applied to masked patches.
- 100 x 4 x 6 x 7 Occluded Images.



01 Evaluation of Inpainted/Occluded Samples

02 Quantitative Evaluation

Evaluation of Inpainted Samples

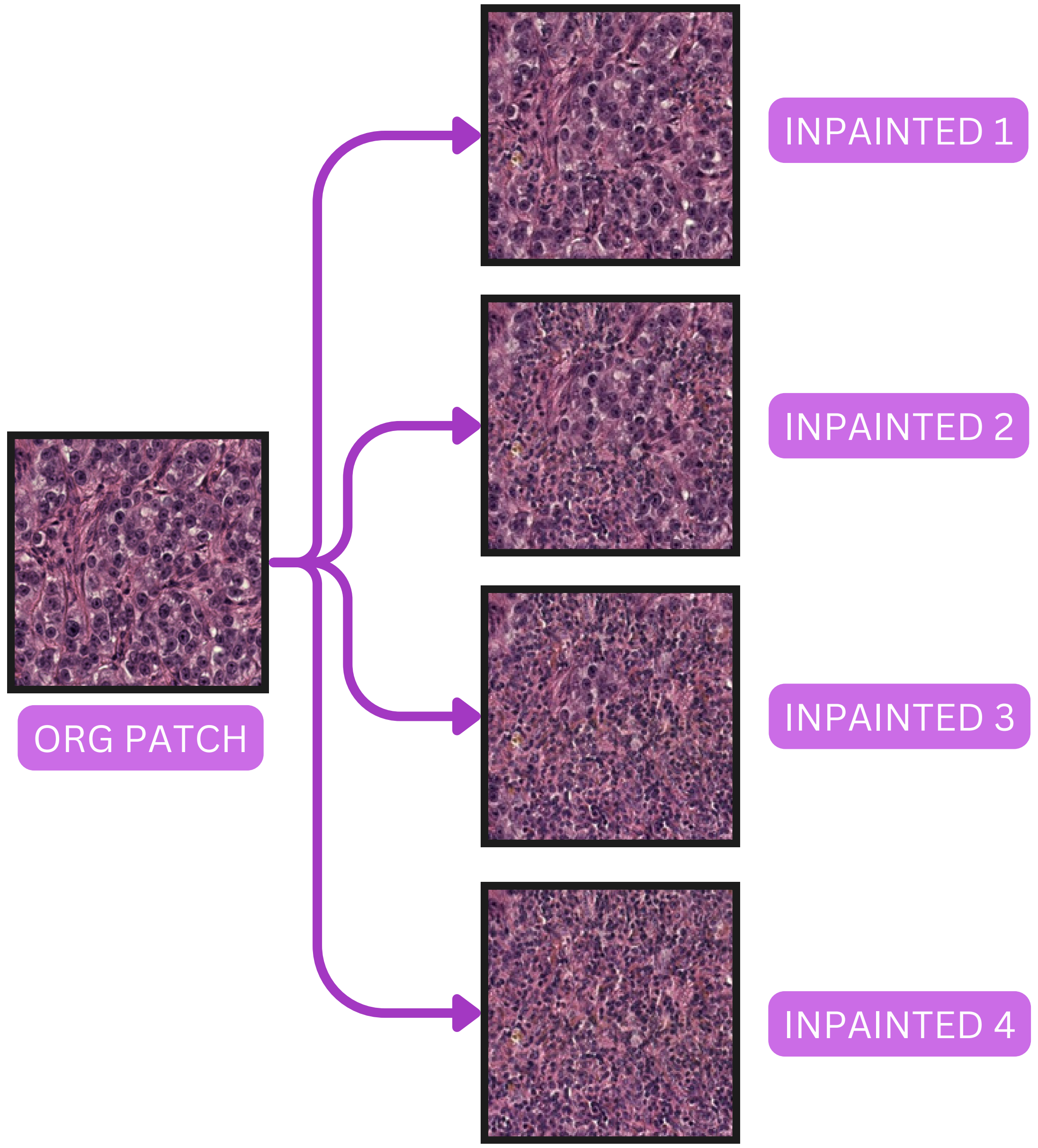
- **LPIPS [12]**

$$\mathrm{LPIPS}(x,\hat{x}) = \sum_{l} \frac{1}{H_l W_l} \sum_{h=1}^{H_l} \sum_{w=1}^{W_l} \left\| \mathbf{w}_l \odot \left(\phi_l(x)_{hw} - \phi_l(\hat{x})_{hw} \right) \right\|_2^2,$$

- **Alexnet**

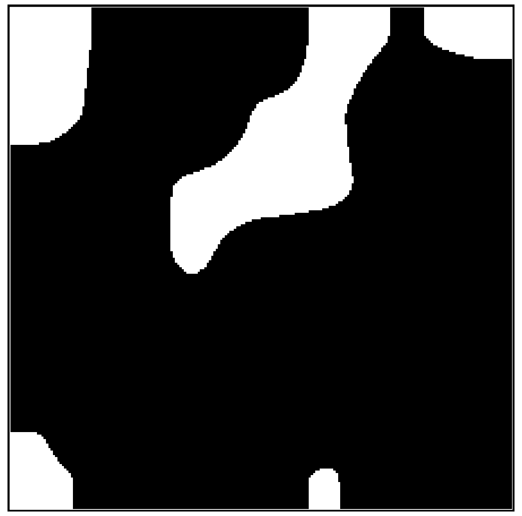
Table 1. LPIPS Scores for different occlusion strategies. Each Mask$_i$ corresponds to the inpainting of specific regions in the image, progressively targeting areas of decreasing importance as identified by the heatmap.

Occlusion Strategy	Mask$_1$	Mask$_2$	Mask$_3$	Mask$_4$
Blackening	0.1106	0.2280	0.3693	0.5567
Histogram	0.0841	0.1844	0.3082	0.4621
Mean	0.0796	0.1776	0.2997	0.4520
NLI	0.0781	0.1769	0.2999	0.4537
Blurring	0.0701	0.1593	0.2670	0.3895
IBO	**0.0381**	**0.0826**	**0.1407**	**0.2180**

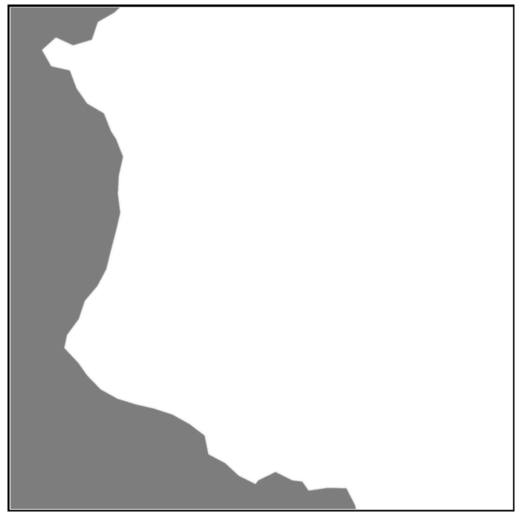


• Quantitative Evaluation

$$\mathrm{IoU} = \frac{R_{\mathrm{ht}} \cap R_{\mathrm{gt}}}{R_{\mathrm{ht}} \cup R_{\mathrm{gt}}}$$



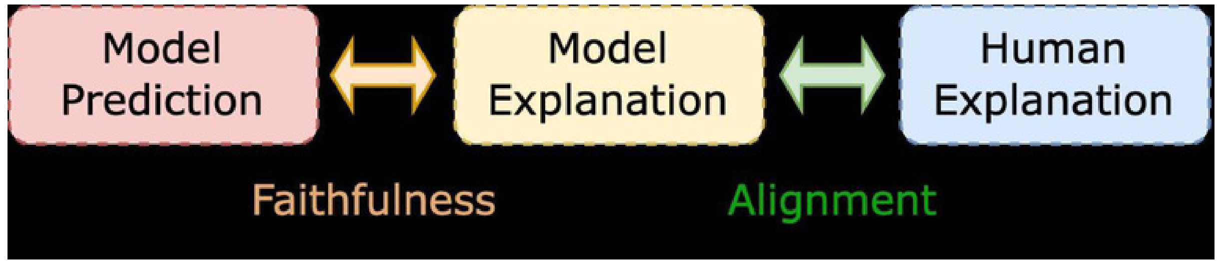
Combined Mask



Ground Truth Mask

Table 2. Mean IoU scores between ground-truth and heatmaps generated by CAM-based approaches across all test samples. This ranking serves as the reference standard for evaluating various occlusion strategies.

Approach	IoU
Full-Grad	0.6896
Grad-CAM	0.6482
Grad-CAM++	0.6467
XGrad-CAM	0.6464
Score-CAM	0.6452
Ablation-CAM	0.6439
Eigen-CAM	0.4257



Results

IoU & AUC

OCCLUSION STRATEGIES RANKINGS (MEAN AUC)

GT RANKINGS (MEAN IOU)

Approach
Full-Grad
Grad-CAM
Grad-CAM++
XGrad-CAM
Score-CAM
Ablation-CAM
Eigen-CAM

$$IoU = \frac{1}{AUC}$$

71%

Table 3. IBO rankings

Approach	AUC
Full-Grad	0.5335
Grad-CAM	0.5991
Grad-CAM++	0.6014
XGrad-CAM	0.6058
Ablation-CAM	0.6363
Score-CAM	0.6707
Eigen-CAM	0.8622

42%

Table 4. Blurring rankings

Approach	AUC
Full-Grad	0.4769
Grad-CAM	0.5360
XGrad-CAM	0.5360
Ablation-CAM	0.5651
Grad-CAM++	0.5898
Score-CAM	0.6094
Eigen-CAM	0.8511

Table 5. NLI rankings

Approach	AUC
Full-Grad	0.3788
Grad-CAM	0.4265
XGrad-CAM	0.4278
Ablation-CAM	0.4612
Grad-CAM++	0.4850
Score-CAM	0.5154
Eigen-CAM	0.7916

Table 6. Histogram rankings

Approach	AUC
Full-Grad	0.3823
Grad-CAM	0.4316
XGrad-CAM	0.4335
Ablation-CAM	0.4634
Grad-CAM++	0.4887
Score-CAM	0.5214
Eigen-CAM	0.7899

Table 7. Mean rankings

Approach	AUC
Full-Grad	0.3718
XGrad-CAM	0.4224
Grad-CAM	0.4233
Ablation-CAM	0.4537
Grad-CAM++	0.4790
Score-CAM	0.5078
Eigen-CAM	0.7871

Table 8. Blackening rankings

Approach	AUC
Full-Grad	0.4857
Grad-CAM++	0.5594
Grad-CAM	0.5624
Score-CAM	0.5897
Ablation-CAM	0.6061
XGrad-CAM	0.6290
Eigen-CAM	0.8630

Mean Absolute Rank Difference (MARD):

$$\text{MARD} = \frac{1}{N}\sum_{i=1}^{N} \left|\text{Rank}_{\text{GT}}(i) - \text{Rank}_{\text{oc}}(i)\right|$$

IoU

Approach
Full-Grad
Grad-CAM
Grad-CAM++
XGrad-CAM
Score-CAM
Ablation-CAM
Eigen-CAM

Table 7. Mean rankings

Approach	AUC
Full-Grad	0.3718
XGrad-CAM	0.4224
Grad-CAM	0.4233
Ablation-CAM	0.4537
Grad-CAM++	0.4790
Score-CAM	0.5078
Eigen-CAM	0.7871

Table 8. Blackening rankings

Approach	AUC
Full-Grad	0.4857
Grad-CAM++	0.5594
Grad-CAM	0.5624
Score-CAM	0.5897
Ablation-CAM	0.6061
XGrad-CAM	0.6290
Eigen-CAM	0.8630

Table 9. MARD values for occlusion strategies.

Occlusion Strategy	MARD Value
Mean	1.1428
Blackening	0.8571
Histogram	0.8571
NLI	0.8571
Blurring	0.8571
IBO	**0.2857**

• Pros:

1. **Better XAI Evaluation:** Reduces Out-of-Distribution (OoD) samples and improves ranking accuracy.
2. **Realistic Inpainting:** Preserves tissue characteristics using DDPM.
3. **Accurate Comparisons:** Closely aligns with ground truth rankings.
4. **Broad Potential:** Framework adaptable for various medical imaging tasks.

• Cons:

1. **High Computation:** Time-intensive and costly for large datasets.
2. **Narrow Scope:** Focused on classification; limited exploration of other applications.
3. **Efficiency Issues:** Needs optimization for scalability and real-time use.

[1] B. Zhou, A. Khosla, L. A., A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. CVPR, 2016.

[2] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision, 128(2):336–359, Oct. 2019.

[3] S. Krishna, S. Suganthi, A. Bhavsar, J. Yesodharan, and S. Krishnamoorthy. An interpretable decision-support model for breast cancer diagnosis using histopathol ogy images. Journal of Pathology Informatics, 14:100319, 2023.

[4] W. Samek, A. Binder, G. Montavon, S. Bach, and K.-R. Müller. Evaluating the visualization of what a deep neural network has learned, 2015.

[5] R. Tomsett, D. Harborne, S. Chakraborty, P. Gurram, and A. Preece. Sanity checks for saliency metrics, 2019.

[6] R. C. Fong and A. Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, Oct. 2017.

[7] P. Wei, Y. Zhang, L. Huang, and Y. Liu. Explainable ai: A survey on approaches and theories. IEEE Transactions on Neural Networks and Learning Systems, 2018.

[8] Y. Rong, T. Leemann, V. Borisov, G. Kasneci, and E. Kasneci. A consis tent and efficient evaluation strategy for attribution methods.

[10] Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA, 318(22):2199–2210, 2017.

[11] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 11461–11471, 2022.

[12] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018.

Sharif University of Technology

Computer Engineering Department

THANK YOU FOR YOUR ATTENTION.