Slide 2: Free Form in painting:

- Free-form inpainting is the task of adding new content to an image in the regions specified by an arbitrary binary mask.
- Most existing approaches train for a certain distribution of masks, which limits their generalization capabilities to unseen mask types.
- Training with pixel-wise and perceptual losses often leads to simple textural extensions towards the missing areas instead of semantically meaningful generation

Slide 3: RePaint

- The paper introduced RePaint: A Denoising Diffusion Probabilistic Model (DDPM) based inpainting approach that is applicable to even extreme masks

- The process is conditioned on the masked input (left). It starts from a random Gaussian noise sample that is iteratively denoised until it produces a high-quality output. Since this process is stochastic, we can sample multiple diverse outputs. The DDPM prior forces a harmonized image, is able to reproduce texture from other regions, and inpaint semantically meaningful content.

Slide 4: Related Works

Past work falls in three groups

- Deterministic Image Inpainting (e.g. DeepFill, LaMa)
    - Good at background extension, repetitive textures.
    - Poor with semantic object completion and large holes.

- Diverse Image Inpainting (e.g. VAE, DSI, ICT)
    - More diverse, but less globally coherent.
    - Autoregressive models are GPU-hungry and struggle with consistency.

- Image Prior & Early Diffusion
    - StyleGAN inversion only works on faces.
    - Early diffusion (Sohl-Dickstein, ILVR, SDEdit) lacked proper evaluation or failed on holes with no frequency info.

Slide 5: RePaint – Key Contributions

Repaint is ready-made diffusion model; we never retrain it for masks.

- Uses an unconditional, pre-trained DDPM — no retraining per mask.
- Introduces conditioning during sampling to fix known pixels.
- Adds resampling schedule to improve harmony at boundaries.
- Works on extreme masks (even when most of the image is missing).
- Produces multiple realistic outputs, not just one.

<u>Slide 6: DDPM</u>

Let me quickly explain what a Denoising Diffusion Probabilistic Model—a DDPM—does.

Start from pure noise

We begin with an image made of random pixels, noted x-T that matches the markovian property xT is totally independent from x0.
every pixel is pure noise drawn from a standard normal distribution N(0, 1).

The forward (noising) process

To train the model we do the opposite: we add a tiny bit of Gaussian noise to a clean image again and again, T times, until it becomes pure noise.
The exact rule is the formula you see as Equation (1):

we keep $\sqrt{(1 – \beta t)}$ of the current image,
we mix in βt of fresh noise.
So as t grows, more noise, less picture.

Train the reverse process

Now we ask a neural network to learn the way backwards.
It looks at a noisy image x_t and predicts the mean $\mu\theta$ and variance $\Sigma\theta$ of a Gaussian that should give us a slightly cleaner image x_{t-1}.
That rule is Equation (2).
During training we compare its guess to the true pair we already built in the forward process.

Sampling—how we generate a picture

Once trained, we flip the chain:

start from pure noise x-T,
apply the learned denoise rule again and again,
after T steps we reach x_0, a sharp, high-quality image.

<u>Slide 7: DDPMs</u>

<u>Slide 8: Repaint Loop</u>

“Each loop of RePaint has two actions:

1. *Keep* the known pixels,
2. *Improve* the unknown pixels.

Input – top-left: the original photo with a purple mask showing the hole.
Add noise to the known part (green ‘noise’ box)

- We copy only the visible pixels and add exactly the amount of Gaussian noise that belongs to the current time-step.
- This gives the top-middle image xt−1~qx_{t-1} \sim qxt−1~q.

Apply the mask (white = hole, black = keep)

- We cut out everything except the noisy visible pixels.
- Result (top-right) holds *only* the correct noisy version of the known region.

Denoise the whole image (orange ‘denoise’ box)

- Bottom-left: xtx_txt is the full noisy picture at this step.
- The DDPM predicts a cleaner version of *all* pixels, noted xt−1~pθ

Mask inverse for the unknown part

- We now keep only the *unknown* pixels from the denoised image (bottom-middle).

Merge the two halves

- Known noisy pixels + unknown denoised pixels → new image xt−1x_{t-1}xt−1 (far right).
- This image is ready for the **next iteration** of the diffusion chain.

<u>Slide 9: Conditioning known regions</u>

<u>Slide 10: Resampling: incoherence, disharmony</u>

<u>Slide 11 :</u>

The algorithm starts with a random noise image.
The loop starts with t = T and runs until t = 1. The algorithm progressively reduces the noise (from xT to a brighter image) by going through several steps.
The inner loop iterates over each pixel u, where each pixel can be:
– Known (already visible part of the image).
– Unknown (missing part to be filled).

Lines 4-5 – redraw the known pixels
If we are not at the final step, make fresh Gaussian noise *c*.
Build a noisy copy of the visible pixels: at time step t-1 using the equation 8a
Lines 6-7 – predict the unknown pixels
Sample another noise z.
Use the DDPM to get the unknown part from the equation 8b
Line 8: fuse both parts
Resampling: line 9-11
If we still have resample loops left (`u < U`) and not at the very end (`t > 1`)
The resampling process is applied, where a small amount of noise is added at each step. This helps generate missing pixels more smoothly with the rest of the image.

Slide 12: Experiments:

Slide 13: Qualitative Results:
Left block = ImageNet examples, *right block* = CelebA-HQ faces.
For each row the first image is the **input with the mask**, then four baselines, and the last column is **RePaint (ours)**."
In every mask type, you can see RePaint keeps the boundary clean and adds semantically correct content, while other methods either blur, show checkerboard artefacts, or hallucinate wrong objects

Slide 14: Metrics:
"We have two measures.
First, a user study: 100 images, six masks, five people, 1 000 votes per pair, with 95 % confidence.
Second, LPIPS: is the metric that computes the similarity between the activations of 2 images for some predefined network which is AlexNet and computes the L2 distance in the feature space, perceptually this metric matches the human perception.

Slide 15: Quantitative Results:
CelebA-HQ (top block)
Human votes: RePaint wins every mask type; the margin is huge on "Alternating Lines" (only 0.7 % prefer ICT vs 99.3 % RePaint).

LPIPS: RePaint is best or second-best on thin/complex masks (Narrow, Alt. Lines), but GAN methods (LaMa) edge it on very "GAN-friendly" masks (Wide, Half).
Take-away: numeric LPIPS does not always track human realism—especially when the inpainted content is semantically different from the ground truth.

ImageNet (bottom block)
Gaps narrow, but RePaint still tops every mask in votes; LaMa is the closest challenger on Wide (42.4 % vs 57.6 %).

LPIPS again shows mixed results: ICT scores better on Wide and Narrow, yet humans still prefer RePaint.
diffusion fills may diverge more from the original photo (hurting LPIPS) but still look more natural to people.

Slide 16: Ablation study

- 32 ImageNet images, **Wide mask** (same ones as LaMa test).
- Slowing down the diffusion process means increasing T, the total number of steps in the denoising chain.
- Adding a few **resample loops** is a much better way to use extra GPU time than simply running more diffusion steps.

Slide17: Main take aways & Limitations

- **Mask-agnostic success** – Same pretrained DDPM works on any mask; wins **all human votes** across 12 tests.
- **Smart resampling** – Extra loops polish the seam and give higher quality for the same compute (see Slide 24).
- **Sharp and diverse** – Generates many realistic options instead of one fixed answer.

Limits right now

- **Speed** – About **10× slower** than GAN or AR methods, so not real-time yet.

- **Metric mismatch** – On very big holes it may look great but score worse on LPIPS because the fill differs from ground truth.

- **Training bias** – Shares the biases of its dataset (e.g., adds too many dogs on ImageNet) and could be misused for fake images.