UAI-NET: UNCERTAINTY-AWARE IMAGE INPAINTING WITH LEARNED ADAPTIVE REFINEMENT DEPTH

Anonymous authors
Paper under double-blind review

ABSTRACT

Most inpainting models fill in missing regions without indicating which parts of the fill are likely to be wrong, and those that do estimate a confidence map use it only to order pixels within a fixed number of refinement passes. This paper makes the model estimate its own difficulty and act on it. UAI-Net predicts a coarse spatial uncertainty map, trained to reproduce the rank ordering of the model's reconstruction error, and uses it twice: to weight the fusion of local, global-attention and boundary features, and to decide, through a straight-through halting head, how many refinement steps (one to four) each location of the refinement grid receives. Training the halting head is the main difficulty: with a continuous depth target its logits collapse to zero and the executed depth is decided by floating-point rounding; an integer target with a small confidence penalty removes the problem. On CelebA-HQ, Places2 and FFHQ (256×256, free-form masks over 10–60% of the image, 1,500 test images each) the map correlates with the true per-pixel error at a within-image Spearman ρ of 0.60–0.67, and the model's predicted variance detects badly filled pixels better than the flip-disagreement or snapshot-ensemble uncertainty of stronger inpainters, in one forward pass. On CelebA-HQ the least-uncertain tercile receives 1.7 refinement steps and the most-uncertain 3.2 (2.54 on average), matching the quality of uniform four-step refinement; the block is still evaluated densely, so wall-clock time is unchanged. Against a PSNR-optimised variant at equal PSNR, UAI-Net gives 8–15% lower LPIPS and 10–17% lower FID.

1 INTRODUCTION

Image inpainting asks a model to fill a masked region consistently with the visible context. Modern approaches based on gated (Yu et al., 2019) and Fourier (Suvorov et al., 2022) convolutions, transformers (Li et al., 2022a) and diffusion (Lugmayr et al., 2022) produce plausible completions but share two limitations. First, most return a single image with no indication of *where* it is likely to be wrong; the exceptions that predict a confidence map (Zeng et al., 2020; Ma et al., 2024) use it to decide which pixels to trust between iterations, and neither measures whether the map predicts error. Second, they apply a fixed number of refinement passes to every masked location, although a pixel at the hole boundary is far easier to complete than one in the middle of a large region of hair or texture.

UAI-Net predicts a coarse (32×32) uncertainty map U at the network bottleneck, supervised to reproduce the *rank order* of the model's own reconstruction error inside the hole, and we verify on 1,500 test images that it does predict error. The map is then used twice. It weights the fusion of a local, a global self-attention and a boundary branch, as in uncertainty-guided fusion for depth completion (Long et al., 2024), so confident regions are completed from local and uncertain regions from global context; and it drives a *learned* per-pixel halting head inside an iterative refinement stage, so that each location of the 128×128 refinement grid receives between one and four refinement steps; unlike the fixed-iteration feedback of Zeng et al. (2020) and Ma et al. (2024), the number of steps a pixel receives is a trained decision with a target, not a heuristic on whether its confidence rose.

Making the halting head trainable is the central technical problem. Training the expected depth toward a continuous target $1+(K-1)U$ produces a model whose halting logits collapse to zero: every

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

stop decision sits exactly on the 0.5 threshold and the executed depth is decided by floating-point rounding (section 3.5). We show this on the logits of a trained model, explain why the continuous target has no minimiser under hard halting, and fix it with an integer target and a confidence penalty.

Contributions.

- An uncertainty map supervised on the *rank* of the model's own reconstruction error, which predicts the true per-pixel error ($\rho = 0.65$–0.67) and detects badly filled pixels better than the multi-pass uncertainty of stronger inpainters.
- A learned per-pixel refinement depth: a straight-through halting head with an integer target whose hard forward pass matches inference, where prior confidence-feedback inpainting fixes the number of rounds.
- The halting-collapse failure mode of continuous depth targets, measured on a trained network's logits, and its fix.

2 Related work

Image inpainting. Partial (Liu et al., 2018) and gated (Yu et al., 2019) convolutions made inpainting networks mask-aware, EdgeConnect (Nazeri et al., 2019) added structure prediction, and larger receptive fields came from Fourier convolutions (Suvorov et al., 2022) and transformers (Li et al., 2022a; Chen et al., 2024; Huang et al., 2024).

Scope. UAI-Net is unconditioned at 256×256 and is compared against unconditioned models with released weights; text-guided (e.g. Manukyan et al., 2025) and zero-shot diffusion inpainting (Lugmayr et al., 2022; Liu et al., 2024a) target prompt-faithful or diverse generation at $\geq 512^2$ and give no per-pixel estimate of where a completion is wrong.

Uncertainty in dense prediction. Heteroscedastic regression (Kendall & Gal, 2017) predicts a per-pixel variance, an auxiliary output here; our routing map is instead trained on the *rank* of the model's own error, robust to the heavy-tailed error inside holes (section 3.2), and controls computation rather than reporting confidence (we claim no calibration, Guo et al. 2017). Pluralistic inpainting (Zheng et al., 2019; Zhao et al., 2020; Cai & Wei, 2020) samples diverse completions but never says where one is wrong.

Confidence-feedback inpainting. Several methods predict a confidence or uncertainty map and feed it back: Zeng et al. (2020) re-generate the pixels whose confidence increased over $T{=}4$ iterations, U^{2}AFN (Ma et al., 2024) treats high-variance pixels as newly missing, SeGuE-Net (Liao et al., 2021) re-estimates poorly inferred regions, and PSM (Li et al., 2024) keeps low-variance predictions and re-predicts the rest (released at 512×512, hence absent from Table 1); SAGI (Giakoumoglou et al., 2025) uses uncertainty in a text-guided pipeline to judge realism. All fix the number of rounds and let the map choose which pixels are updated; the "stop" of Zeng et al. (2020) is a heuristic, and none checks whether its map predicts reconstruction error. UAI-Net differs on all three: its map is supervised on the rank of the model's own error and evaluated against it (§3.2), depth is a learned policy with an explicit target (§3.4), and the same map routes features, a single-map variant of the uncertainty-guided fusion of Long et al. (2024), for which we claim no novelty.

Adaptive computation. Adaptive Computation Time (Graves, 2016) and its spatial variant SACT (Figurnov et al., 2017) learn how many times to apply a weight-shared block; PonderNet (Banino et al., 2021) trains the expected output under a halting distribution. Our refinement stage follows this family with a hard forward and straight-through backward pass (Bengio et al., 2013). FreqExit (Li et al., 2025) learns early exits for visual autoregressive generation, but they are per-sample decisions about model depth, not a spatially varying number of refinement updates. The logit collapse of section 3.5 is, to our knowledge, unreported in this literature, likely because ACT-style methods regularise total ponder cost rather than target a per-pixel depth.

3 Method

Figure 1 shows the network. Input is the damaged image $I \odot (1{-}M)$ concatenated with the mask M; output is $\hat{I} = (1{-}M) \odot I + M \odot \mathrm{rgb}$, so known pixels are reproduced exactly.

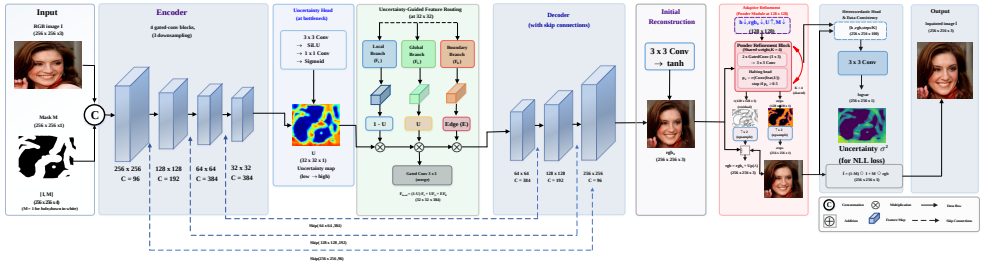


Figure 1: **UAI-Net.** The uncertainty map U (32×32), trained on the rank of the model's own reconstruction error, weights the fusion of the local, global and boundary branches (§3.3) and drives the refinement stage (§3.4): one weight-shared block at 128×128, applied up to $K{=}4$ times, with a halting head deciding per cell when to stop. A heteroscedastic head predicts the log-variance and the output is composited with the known pixels.

3.1 Backbone

A four-stage gated-convolution encoder (Yu et al., 2019) maps the $256{\times}256{\times}4$ input to a $32{\times}32{\times}4c$ bottleneck f ($c{=}96$). Three branches operate on f: *local* (two dilated residual blocks), *global* (4-head self-attention over the 32×32 grid and a dilation-4 residual block) and *boundary* (a residual block and a 3×3 convolution), the last also emitting an edge gate $E \in [0,1]^{32\times32}$. A three-stage decoder with skips produces $h \in \mathbb{R}^{256\times256\times c}$ and an initial estimate $\mathrm{rgb}_0 = \tanh(\mathrm{conv}(h))$.

3.2 Coarse uncertainty head

A head on the bottleneck, $U = \sigma(\mathrm{conv}_{1\times1}(\mathrm{SiLU}(\mathrm{conv}_{3\times3}(f))))$, predicts a 32×32 uncertainty map. It is supervised by the model's own error: let $e = \|\hat{I} - I\|_1$ averaged over channels and average-pooled to 32×32, restricted to cells that touch the hole. The target is the *per-image rank* of e among hole cells, normalised to $[0,1]$, and the loss is

$$\mathcal{L}_{\mathrm{unc}} = \frac{\sum_{x\in\mathcal{H}} \left|U(x) - \mathrm{rank}_{\mathcal{H}}(e)(x)\right|}{|\mathcal{H}|}, \tag{1}$$

where $\mathcal{H}$ is the set of hole cells. Rank normalisation matters. The error inside a hole is heavy-tailed; normalising by the per-image maximum, as one might first try, places a single cell near 1 and compresses the rest into the bottom third of the range, and the downstream depth target $1{+}(K{-}1)U$ then never asks for deep refinement. With ranks, U is uniform on $[0,1]$ over the hole by construction, and the loss is restricted to hole cells so the objective is not dominated by the trivial "predict zero outside the hole".

3.3 Uncertainty-routed feature fusion

The three branch outputs are fused as

$$F_{\mathrm{fused}} = (1{-}U)\,F_L + U\,F_G + E\,F_B, \tag{2}$$

followed by a gated 3×3 merge. Confident cells are completed from local context; uncertain cells draw on global attention; the boundary term is gated by the predicted edge map.

3.4 Adaptive refinement depth

Refinement runs at 128×128: the decoder feature, the initial estimate, the mask and U are bilinearly resampled to that grid once, and we drop the resampling symbols below. It operates on $[h, \mathrm{rgb}_0, U]$ with a single weight-shared block $\mathcal{R}$ (two gated convolutions and a 3×3 convolution to an RGB residual). Let $a_t \in \{0,1\}$ mark cells of the 128×128 refinement grid still active at step t (initially the hole); "per-pixel" depth below always refers to this grid, whose step map is upsampled to the output; Figure 2 shows the loop on a real test image. One step is

$$\mathrm{feat}_t = \mathcal{R}([h_t, \mathrm{rgb}_t, U]), \qquad h_{t+1} = h_t + a_t \odot \mathrm{feat}_t, \tag{3}$$

$$\mathrm{rgb}_{t+1} = a_t \odot \tanh(\mathrm{rgb}_t + \mathrm{conv}(\mathrm{feat}_t)) + (1{-}a_t) \odot \mathrm{rgb}_t, \tag{4}$$

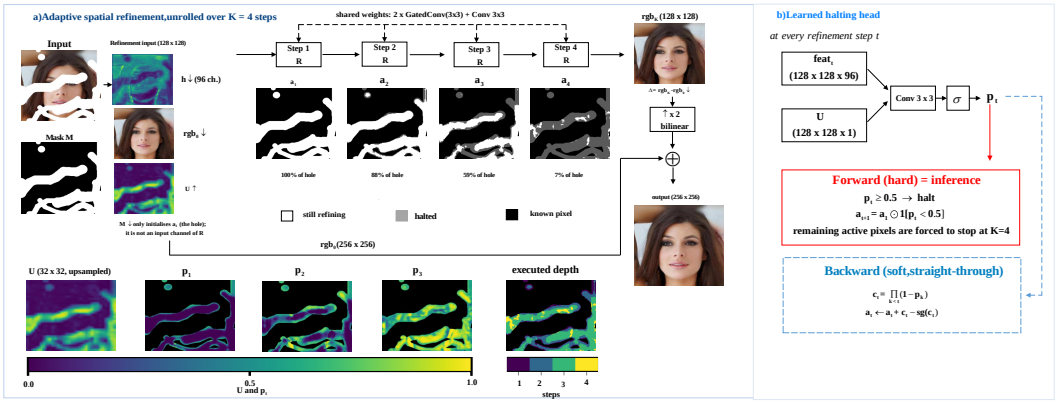


Figure 2: **Adaptive refinement, on real outputs of the trained model.** *(a)* Only cells that have not halted are updated; on this CelebA-HQ image (30–40% mask) 100%, 88%, 59% and 7% of the hole is active at steps 1–4. *(b)* The forward pass is hard and identical to inference ($p_t \geq 0.5$ halts); the backward pass is straight-through. The executed depth is decided by p_t, not computed from U.

and a halting head $p_t = \sigma(\mathrm{conv}_{3\times3}([\mathrm{feat}_t, U]))$ decides who stops: $a_{t+1} = a_t \odot \mathbf{1}[p_t < 0.5]$, with all remaining pixels forced to halt at $t{=}K$. Each pixel's output is frozen at the step it halts. The residual $\mathrm{rgb}_K - \mathrm{rgb}_0$ is upsampled to $256{\times}256$ and added to rgb_0.

Straight-through training. Algorithm 1 states the loop as executed; p_t is distinct from the predictive variance σ^2 of §3.6. For the backward pass we substitute the soft continuation probability $c_t = \prod_{k<t}(1{-}p_k)$ through a straight-through estimator, $a_t \leftarrow a_t + c_t - \mathrm{sg}(c_t)$, so that gradients reach the halting head for every step, including steps after a pixel has halted. A hard-only mask zeroes those gradients and the head cannot learn.

Depth loss. The expected depth $\bar{s} = \sum_t c_t$ is trained toward an *integer* target,

$$\mathcal{L}_{\mathrm{depth}} = \mathrm{SmoothL1}\big(\bar{s},\ \mathrm{round}(1 + (K{-}1)U)\big), \tag{5}$$

a target a hard forward pass can actually reach, and a confidence penalty keeps the halting probabilities away from the threshold,

$$\mathcal{L}_{\mathrm{conf}} = \frac{1}{|\mathcal{H}|(K{-}1)} \sum_{t<K} \sum_{x\in\mathcal{H}} p_t(1{-}p_t). \tag{6}$$

averaged over hole pixels. The confidence term pushes p_t toward $\{0, 1\}$. Why the target is rounded is the subject of the next section.

3.5 Halting collapse

Our first version trained $\bar{s}$ toward the continuous target $1 + (K{-}1)U$. It trained without incident: the reported depth tracked its target and the loss was flat, but the model was broken. On the final checkpoint of that run (60 epochs), 100.0% of the 0.9M halting logits over the hole cells of 50 test images lay within ±0.01 of zero (median $|\mathrm{logit}| = 0.0005$; Figure 15), i.e. every pixel's stop probability was 0.5 to three decimal places; after 30 epochs the fraction was already 99.9%. At that point the model executed 3.7 refinement steps per pixel when evaluated in fp32 and 2.7 when evaluated in fp16 (as trained): the depth was being decided by rounding.

The cause is that under a hard forward pass a pixel's depth is an integer in $\{1, \ldots, K\}$, while its target, e.g. 2.5, is not. The loss cannot reach zero, and its gradient pushes the decisive logit toward the threshold from both sides (a pixel at depth 2 is told to go deeper, at depth 3 to go shallower), so the head learns to sit exactly on it. Rounding the target gives every pixel a reachable minimiser; the confidence penalty removes the residual soft solutions. With this change the median logit margin grows monotonically during training (0.32 after one epoch of a short run to 0.62 after four) instead of shrinking, reaching 1.56 in the final UAI-Net model, where only 0.4% of logits lie within ±0.01; the tercile depth spread widens from 0.4 to 1.5 steps. Figure 15 shows both distributions. We report this because it is easy to miss: nothing in the training log signals it.

Algorithm 1 Adaptive refinement with hard halting (inference; training differs only in the backward pass)

Require: decoder feature h, initial estimate rgb_0, uncertainty U (32×32), hole mask M, steps $K{=}4$

$h_1, \mathrm{rgb}_1, a_1 \leftarrow h, \mathrm{rgb}_0, M$
for $t = 1$ **to** K **do**
  $\mathrm{feat}_t \leftarrow \mathcal{R}([h_t, \mathrm{rgb}_t, U])$
  $h_{t+1} \leftarrow h_t + a_t \odot \mathrm{feat}_t$
  $\mathrm{rgb}_{t+1} \leftarrow a_t \odot \tanh(\mathrm{rgb}_t + \mathrm{conv}(\mathrm{feat}_t)) + (1 - a_t) \odot \mathrm{rgb}_t$
  $p_t \leftarrow \sigma(\mathrm{conv}_{3\times3}([\mathrm{feat}_t, U]))$
  $a_{t+1} \leftarrow a_t \odot \mathbf{1}[p_t < 0.5]$
  if $\sum a_{t+1} = 0$ **then break**
  end if
end for
$\mathrm{rgb} \leftarrow \mathrm{rgb}_0 + \mathrm{Up}(\mathrm{rgb}_{t+1} - \mathrm{rgb}_0)$
return $(1 - M) \odot I + M \odot \mathrm{rgb}$

3.6 Losses

The heteroscedastic head predicts a per-pixel log-variance $\ell = \mathrm{conv}_{3\times3}([h, \mathrm{rgb}, s/K])$ from the refined output and the step map, trained with the Gaussian negative log-likelihood $\mathcal{L}_{\mathrm{NLL}} = \frac{1}{2}(e^{-\ell}\,\|\hat{I}-I\|^2 + \ell)$ over the hole (Kendall & Gal, 2017). The full generator objective is

$$\begin{aligned}\mathcal{L}_G = \lambda_{\mathrm{rec}}\mathcal{L}_1 + \lambda_{\mathrm{nll}}\mathcal{L}_{\mathrm{NLL}} + \lambda_{\mathrm{unc}}\mathcal{L}_{\mathrm{unc}} + \lambda_{\mathrm{dep}}\mathcal{L}_{\mathrm{depth}} + \lambda_{\mathrm{conf}}\mathcal{L}_{\mathrm{conf}} \\ + \lambda_{\mathrm{perc}}\mathcal{L}_{\mathrm{perc}} + \lambda_{\mathrm{sty}}\mathcal{L}_{\mathrm{sty}} + \lambda_{\mathrm{adv}}\mathcal{L}_{\mathrm{adv}} + \lambda_{\mathrm{fm}}\mathcal{L}_{\mathrm{fm}},\end{aligned} \tag{7}$$

with weights $3, 1, 2, 0.05, 0.02, 2, 120, 0.75, 10$. The reconstruction, perceptual, style, adversarial and feature-matching terms are the standard ones of Suvorov et al. (2022); Appendix F gives each term and its schedule.

4 Experiments

Setup. All three datasets are used at 256×256 with free-form masks (Yu et al., 2019) whose hole ratio is drawn from 10–60%, fixed across all runs: CelebA-HQ (Karras et al., 2018) (28,000 training / 1,500 test images, 50 epochs), Places2 (Zhou et al., 2017) (50,000 of $\sim$1.8M / 1,500 validation images, 30 epochs) and FFHQ (Karras et al., 2019) (50,000 / 1,500, 30 epochs), which is $\sim$187k iterations against CelebA-HQ's $\sim$175k. The model has 32.4M parameters ($c{=}96$), $K{=}4$, and trains on one RTX 4090 in $\sim$12 h; checkpoints are selected by LPIPS, since PSNR selection prefers blurrier models. We report PSNR, SSIM, LPIPS (Zhang et al., 2018) and FID (Heusel et al., 2017) per bucket on all 1,500 test images, FID against the ground truth of the same bucket; every number is from this protocol unless stated (Appendix F).

Baselines. DeepFillv2 (Yu et al., 2019), LaMa (Suvorov et al., 2022), MISF (Li et al., 2022b), the pluralistic latent-code model of Chen & Zhao (2024), HINT (Chen et al., 2024) and SPA-Former (Huang et al., 2024) run with their official weights on exactly our images, masks and metric code (†: no FFHQ release, CelebA-HQ weights applied); StrDiffusion (Liu et al., 2024b) (72 s per image) is evaluated the same way on a 200-image matched subset (Table 6). Published numbers use other masks and splits and are not comparable. Every method sees the same input and has its known pixels restored before scoring, as UAI-Net does.

4.1 Main results

Table 1 gives the two extreme buckets and Figure 3 two representative completions; the full five-bucket tables are in Appendix A. The two transformer models are more accurate under the identical protocol: SPA-Former and HINT lead by 0.9–1.6 dB on CelebA-HQ, HINT by 1.3–1.8 dB on Places2 (where UAI-Net saw 50,000 of $\sim$1.8M images), their CelebA-HQ weights by 0.6–1.1 dB on FFHQ. UAI-Net leads DeepFillv2 and LaMa on the face datasets, is within 0.3 dB of MISF there, 0.2–1.0 dB behind the latent-code model on CelebA-HQ but 1.2–1.7 dB ahead on Places2, and

Table 1: **Fill quality**: 1,500 test images per bucket, identical fixed masks and metric code. Baselines use their official weights on our images and masks (†: no FFHQ release, CelebA-HQ weights applied). "PSNR-opt." and "+ more data" are variants of UAI-Net (Appendix F).

		10–20%				50–60%			
Dataset	Method	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
CelebA-HQ	DeepFillv2	30.78	0.953	0.0314	6.03	22.80	0.799	0.1399	29.28
	LaMa	30.72	0.954	0.0216	4.18	22.45	0.795	0.1081	16.02
	MISF	31.54	0.957	0.0200	3.86	23.19	0.807	0.0949	13.60
	Latent codes	31.75	0.956	0.0191	3.83	23.79	0.814	0.0839	12.51
	HINT	32.50	0.962	0.0158	3.23	24.14	0.829	**0.0751**	12.08
	SPA-Former	**32.85**	**0.964**	**0.0153**	**3.04**	**24.48**	**0.838**	0.0753	**10.73**
	UAI-Net, PSNR-opt.	31.52	0.959	0.0247	4.78	22.96	0.811	0.1108	18.52
	UAI-Net	31.55	0.958	0.0211	4.14	22.90	0.804	0.1020	15.34
Places2	DeepFillv2	27.06	0.928	0.0481	9.46	19.34	0.695	0.2064	36.71
	LaMa	27.86	0.932	0.0350	6.60	20.35	0.718	0.1569	22.71
	MISF	28.02	0.932	0.0388	7.47	20.27	0.712	0.1714	27.08
	Latent codes	26.03	0.920	0.0457	8.24	18.05	0.669	0.2074	27.92
	HINT	**29.01**	**0.941**	**0.0307**	**5.61**	**21.06**	**0.740**	**0.1412**	**20.32**
	SPA-Former	28.74	0.939	0.0318	5.71	20.90	0.735	0.1490	20.98
	UAI-Net	27.73	0.931	0.0423	8.16	19.30	0.698	0.1909	30.40
	UAI-Net + more data	27.96	0.934	0.0398	7.59	19.34	0.704	0.1816	27.70
FFHQ	DeepFillv2†	29.74	0.949	0.0346	8.95	21.84	0.779	0.1593	40.88
	LaMa	29.86	0.950	0.0233	5.78	21.73	0.783	0.1184	21.44
	MISF†	30.35	0.952	0.0237	6.03	22.13	0.787	0.1133	24.27
	HINT†	**31.37**	**0.959**	**0.0175**	**4.48**	**22.94**	**0.813**	**0.0866**	**15.13**
	SPA-Former†	31.24	0.958	0.0194	4.99	22.91	0.813	0.0955	18.99
	UAI-Net	30.66	0.956	0.0223	5.82	21.89	0.790	0.1117	21.07

ahead of StrDiffusion on a matched subset (Table 6). We do not claim state-of-the-art fidelity; what is evaluated below is the uncertainty map and the allocation it drives. (HINT's and SPA-Former's CelebA-HQ weights come from the authors' own splits, which we cannot verify to be disjoint from our test images; Places2 uses the official validation set.) At equal PSNR the perceptually-weighted UAI-Net improves LPIPS by 8–15% and FID by 10–17% over the PSNR-optimised variant. Fine-tuning Places2 for six epochs on 300,000 images raises PSNR by 0.04–0.23 dB and lowers FID by 7–9% in every bucket, leaves Table 2 unchanged within its intervals, and does not close the gap to HINT: the remaining difference is architectural.

4.2 Uncertainty versus error

For every test image we compute the Spearman rank correlation between U (upsampled) and the absolute reconstruction error over hole pixels. The mean within-image ρ is 0.669, 0.674, 0.672, 0.664 and 0.651 for the five buckets (Pearson 0.43–0.44). Figure 4 shows what this looks like: U is high on hair, eyes and mouths inside the hole and low on skin and background, and the error map follows it. The map is trained only on the rank of the model's own error and receives no ground-truth uncertainty. Figure 14 shows the per-step halting probabilities that U drives, and Figure 9 the images where the model is most wrong.

4.3 Error detection

Rank correlation says the map orders pixels roughly right; a user needs to know whether it *finds* the pixels a completion got wrong. We treat every uncertainty source as a detector of the 20% of hole pixels with the largest error and report ROC and precision–recall areas, the within-image ρ, the sparsification error (AUSE) and its gain over random removal (AURG) (Ilg et al., 2018), and the hole-PSNR gain when the 10% most uncertain hole pixels are replaced by the truth (random: $+0.46$ dB). Sources for UAI-Net are its variance head $\sigma^2 = e^{\ell}$, the map U, flip disagreement and the disagreement of the last three checkpoints (a snapshot ensemble, Lakshminarayanan et al. 2017);

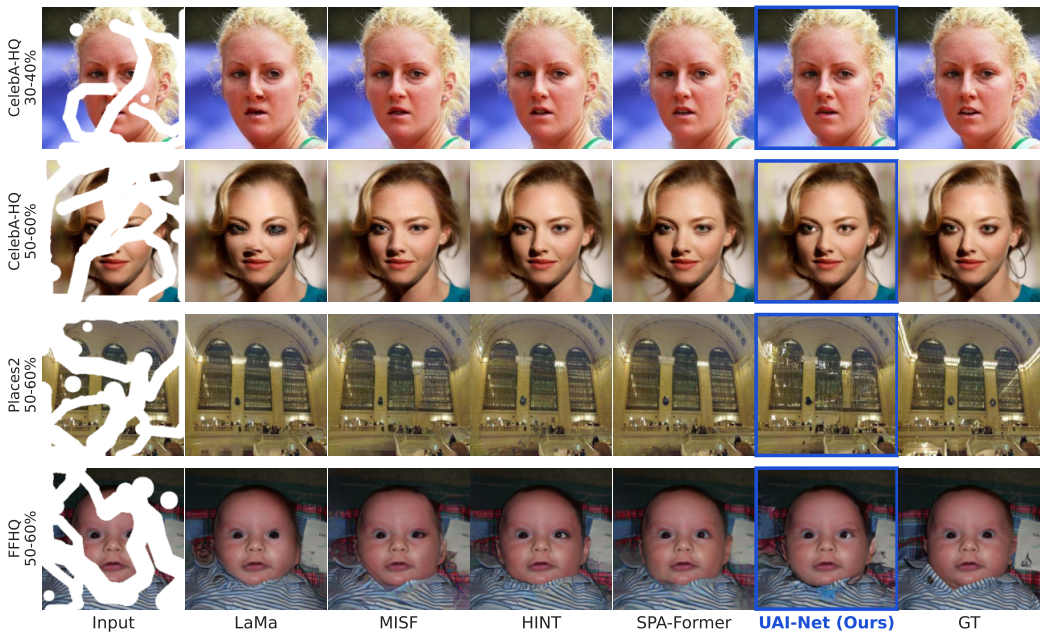


Figure 3: **Qualitative comparison across the three datasets**; hole in white, our column boxed. Each row is the median-PSNR image of that bucket, not a selected example. DeepFillv2 and the latent-code model, and every mask bucket, are in Appendix C.

LaMa, HINT and SPA-Former predict no uncertainty, so we use their flip disagreement, with distance to the nearest known pixel as a training-free reference. U^{2}AFN (Ma et al., 2024), Zeng et al. (2020) (no weights at 256×256) and PSM (Li et al., 2024) (512×512) are not included, and none of them reports error-detection metrics. All numbers use 500 images per bucket with 95% bootstrap intervals.

Table 2 reports the averages. The variance head is the best detector on every dataset and metric: AUROC 0.875 / 0.858 / 0.874 on CelebA-HQ / Places2 / FFHQ against 0.761–0.803 for the flip disagreement of LaMa, HINT and SPA-Former and 0.618–0.685 for the distance reference, its bootstrap interval lying entirely above every other source in all three datasets and all five buckets (Table 9). It roughly halves the sparsification error of the strongest baseline detector (AUSE 0.132–0.165 vs. 0.251–0.312) and doubles the value of a fixed correction budget (Fix-10%: $+3.11$ dB on CelebA-HQ vs. $+2.03$ dB for HINT's flip disagreement). One forward pass also beats UAI-Net's own multi-pass alternatives, flip disagreement ($2\times$ compute) and the snapshot ensemble ($3\times$). The routing map U ranks pixels well (ρ_{img} 0.607–0.662, second only to σ^2) but, being 32×32 and meant to allocate depth, detects individual bad pixels less sharply; the baseline detectors order by fidelity (HINT's beats LaMa's), so the advantage of σ^2 is not an artefact of UAI-Net being easier to predict. Figures 11 and 12 draw the sparsification, ROC and reliability curves behind these numbers. The calibration error of σ^2 (UCE over 15 equal-count variance bins: mean absolute difference between a bin's mean predicted variance and its mean squared error, Laves et al. 2020) is $2.2\text{–}5.5\times10^{-4}$ in $[0,1]$ intensity units.

4.4 Refinement depth

Table 3 compares depth policies, which all share one set of trained weights; Figure 13 shows them on the same test images. Learned halting assigns 1.7 steps to the least-uncertain tercile and 3.2 to the most, 2.5 on average, and is better than running all four steps everywhere in all fifteen bucket–metric cells (Table 7). Against the schedule $\text{round}(1{+}3U)$ it was trained toward it reaches comparable quality (better in PSNR and LPIPS in all five buckets, worse in FID in three) with fewer steps (2.54 vs 2.70) and a wider depth range. Does refinement help at all? Barely: a model retrained with a

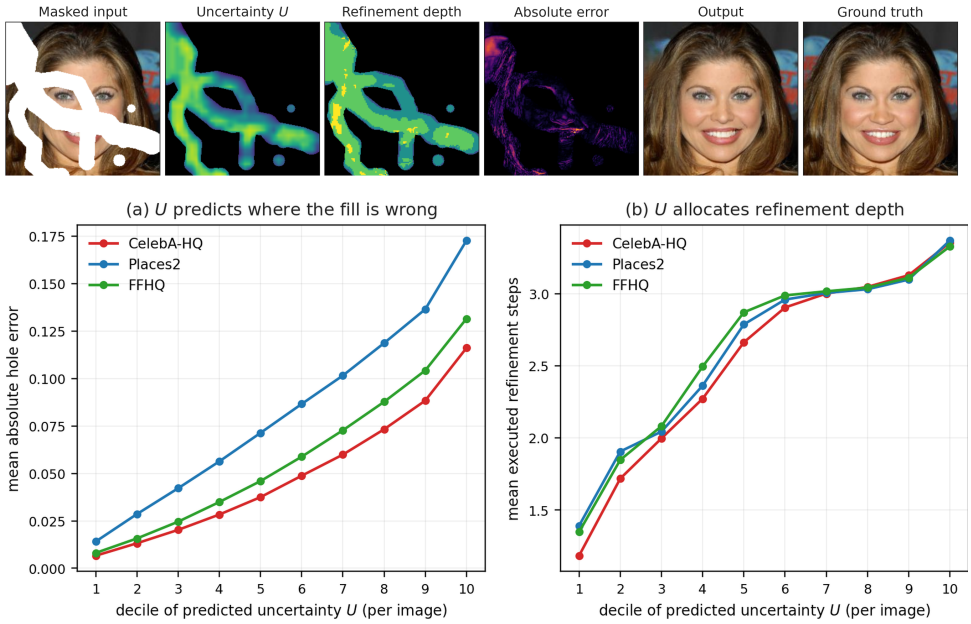


Figure 4: **Uncertainty predicts error and sets refinement depth.** *Top:* one CelebA-HQ image at 30–40% masking; known pixels are black in the maps. *Bottom:* over 300 images per dataset, *(a)* the mean hole error rises with the decile of U by 12.3–17.5$\times$ and *(b)* the executed depth from 1.2–1.4 to $\approx$3.35 steps, a learned relation, since p_t decides the depth.

single step is within 0.07 dB PSNR, 0.7% LPIPS and 1.3% FID of the full model in all five buckets, and its uncertainty map is unaffected ($\rho_{\text{img}} = 0.65$–0.68). We therefore do not claim that adaptive depth improves quality: the experiments establish that a halting head can allocate refinement by uncertainty without the collapse of §3.5, and that its 2.5-step allocation loses nothing against uniform four-step refinement, a saving realisable only under sparse execution (§5). The no-uncertainty row removes routing and the U supervision together and halts after one step, so it is effectively the plain backbone; against it the full mechanism is worth $+0.06$ to $+0.09$ dB and 0.8–1.4% LPIPS in all five buckets and 1.1–1.7% FID in four. Because the gaps are small we test them on the same images and masks: paired Wilcoxon tests give $p < 10^{-5}$ (PSNR) and $p < 0.01$ (LPIPS) for the full model over no-uncertainty in every bucket, $p < 10^{-3}$ (PSNR, four buckets) and $p < 0.01$ (LPIPS, all five) over $K{=}1$, and significance in seven of ten cells for $K{=}1$ over no-uncertainty, a pair that keeps supervision and routing but fixes the depth and so measures the two together (Appendix B). What stands at full strength is the uncertainty map, which reaches $\rho_{\text{img}} \approx 0.67$ however it is used, and the analysis of how to train the halting head without collapse.

5 Limitations

(i) No wall-clock saving is realised (303.9 GFLOPs, 11.9 ms per image, RTX 4090): the dense block still executes all K iterations, and exploiting the 36% fewer per-pixel updates needs sparse execution (Ren et al., 2018). (ii) U is evaluated as an error-ranking signal, not a predictor of error magnitude. (iii) HINT and SPA-Former remain 0.9–1.8 dB ahead under the identical protocol, though the mechanisms do not depend on the backbone. (iv) Experiments are at 256$\times$256, one training run per configuration, Places2 on 50,000 of $\sim$1.8M images (Appendix F).

6 Conclusion

UAI-Net makes an inpainting network estimate where it is likely to fail and act on it twice: to choose between local and global features, and to set each location's refinement depth. The map predicts

Table 2: **Error detection by uncertainty**, averaged over the five buckets (500 images each; per-bucket intervals in Table 9). "Bad": the 20% of hole pixels with the largest error of that method. Fix-10%: hole-PSNR gain when those pixels are corrected (random: $+0.46$ dB).

Dataset	Uncertainty source	AUROC↑	AUPR↑	ρ_{img} ↑	AUSE↓	AURG↑	Fix-10%↑
CelebA-HQ	UAI-Net variance head σ^2	**0.875**	**0.594**	**0.709**	**0.132**	**0.571**	**+3.11**
	UAI-Net routing map U	0.814	0.472	0.662	0.177	0.526	+1.94
	UAI-Net flip disagreement	0.804	0.512	0.591	0.234	0.469	+2.02
	UAI-Net snapshot ensemble	0.824	0.517	0.622	0.202	0.502	+2.03
	LaMa flip disagreement	0.767	0.469	0.504	0.300	0.392	+1.80
	HINT flip disagreement	0.798	0.505	0.562	0.251	0.446	+2.03
	SPA-Former flip disagreement	0.784	0.490	0.524	0.274	0.417	+1.99
	distance to known pixels	0.671	0.311	0.354	0.374	0.329	+1.04
Places2	UAI-Net variance head σ^2	**0.858**	**0.554**	**0.658**	**0.165**	**0.537**	**+2.43**
	UAI-Net routing map U	0.763	0.401	0.607	0.210	0.492	+1.65
	UAI-Net flip disagreement	0.791	0.477	0.552	0.269	0.433	+1.70
	UAI-Net snapshot ensemble	0.803	0.473	0.580	0.246	0.456	+1.58
	LaMa flip disagreement	0.773	0.461	0.494	0.310	0.379	+1.58
	HINT flip disagreement	0.803	0.487	0.570	0.257	0.453	+1.67
	SPA-Former flip disagreement	0.784	0.468	0.520	0.289	0.407	+1.61
	distance to known pixels	0.618	0.263	0.270	0.454	0.247	+0.70
FFHQ	UAI-Net variance head σ^2	**0.874**	**0.591**	**0.699**	**0.140**	**0.567**	**+2.81**
	UAI-Net routing map U	0.806	0.464	0.653	0.188	0.519	+1.76
	UAI-Net flip disagreement	0.810	0.520	0.597	0.238	0.469	+1.90
	UAI-Net snapshot ensemble	0.835	0.543	0.639	0.199	0.507	+1.95
	LaMa flip disagreement	0.761	0.459	0.500	0.312	0.375	+1.55
	distance to known pixels	0.685	0.327	0.378	0.358	0.349	+1.04

Table 3: **Refinement-depth ablation** on CelebA-HQ (1,500 images). $K{=}1$ and no-U are retrained with identical settings; "steps" is the mean number of refinement updates per hole cell.

			10–20%			50–60%		
Depth policy	Steps	Tercile lo/hi	PSNR	LPIPS	FID	PSNR	LPIPS	FID
Uniform, all 4 steps	4.00	4.0 / 4.0	31.50	0.0215	4.32	22.84	0.1033	15.94
Scheduled $\text{round}(1{+}3U)$	2.70	2.0 / 3.3	31.54	0.0211	4.15	22.88	0.1022	15.27
Learned halting (ours)	**2.54**	1.7 / 3.2	**31.55**	**0.0211**	**4.14**	**22.90**	**0.1020**	**15.34**
$K{=}1$ (single step, retrained)	1.00	1.0 / 1.0	31.53	0.0212	4.18	22.83	0.1027	15.51
No uncertainty (routing off, U untrained, retrained)	1.03	–	31.49	0.0213	4.21	22.82	0.1034	15.59

error, the learned allocation matches uniform refinement at 2.54 steps per cell on CelebA-HQ, and we document a halting-head failure that is silent in the loss and visible only in the logits.

AI USE STATEMENT

Generative AI tools (large language models) were used in this work for: implementing and debugging the training and evaluation code, including the diagnosis of the halting-logit collapse described in §3.5; designing experiments and ablations; and drafting and editing the text of this paper. They were not used to generate data, to prove mathematical claims, or to produce any reported number. All code produced with AI assistance was run, inspected and tested by the authors; every quantitative result was obtained from the authors' own training runs and evaluated on the authors' fixed test protocol; every figure showing model outputs was rendered from a trained checkpoint. The authors reviewed all AI-assisted text and take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

All experiments use a single script with a fixed configuration; the exact training and evaluation commands, the mask-generation procedure with its fixed seed, and the per-bucket test protocol are given in Appendix F. Code and the fixed test masks will be released. Every number in the paper is

from the 1,500-image protocol described in §4, except the error-detection metrics (500 images per bucket, stated in §4.3) and the StrDiffusion comparison, which uses the explicitly stated 200-image matched subset reported in Table 6.

REFERENCES

Andrea Banino, Jan Balaguer, and Charles Blundell. PonderNet: Learning to ponder. In *ICML Workshop on Automated Machine Learning*, 2021.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Weiwei Cai and Zhanguo Wei. PiiGAN: Generative adversarial networks for pluralistic image inpainting. *IEEE Access*, 8, 2020.

Haiwei Chen and Yajie Zhao. Don't look into the dark: Latent codes for pluralistic image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7591–7600, 2024.

Shuang Chen, Amir Atapour-Abarghouei, and Hubert P. H. Shum. HINT: High-quality INpainting transformer with mask-aware encoding and enhanced attention. *IEEE Transactions on Multimedia*, 2024.

Michael Figurnov, Maxwell D Collins, Yukun Zhu, Li Zhang, Jonathan Huang, Dmitry Vetrov, and Ruslan Salakhutdinov. Spatially adaptive computation time for residual networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1039–1048, 2017.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423, 2016.

Paschalis Giakoumoglou, Dimitrios Karageorgiou, Symeon Papadopoulos, and Panagiotis C. Petrantonakis. SAGI: Semantically aligned and uncertainty guided AI image inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Alex Graves. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*, 2016.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, pp. 1321–1330, 2017.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Wenli Huang, Ye Deng, Siqi Hui, Yang Wu, Sanping Zhou, and Jinjun Wang. Sparse self-attention transformer for image inpainting. *Pattern Recognition*, 145:109897, 2024.

Eddy Ilg, Özgün Çiçek, Silvio Galesso, Aaron Klein, Osama Makansi, Frank Hutter, and Thomas Brox. Uncertainty estimates and multi-hypotheses networks for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 694–711, 2016.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4401–4410, 2019.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Max-Heinrich Laves, Sontje Ihler, Jacob F. Fast, Lüder A. Kahrs, and Tobias Ortmaier. Well-calibrated regression uncertainty in medical imaging with deep learning. In *Medical Imaging with Deep Learning (MIDL)*, 2020.

Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. MAT: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10758–10768, 2022a.

Wenbo Li, Xin Yu, Kun Zhou, Yibing Song, Zhe Lin, and Jiaya Jia. Image inpainting via iteratively decoupled probabilistic modeling. In *International Conference on Learning Representations (ICLR)*, 2024.

Xiaoguang Li, Qing Guo, Di Lin, Ping Li, Wei Feng, and Song Wang. MISF: Multi-level interactive Siamese filtering for high-fidelity image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1869–1878, 2022b.

Ying Li, Chengfei Lv, and Huan Wang. FreqExit: Enabling early-exit inference for visual autoregressive models via frequency-aware guidance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Liang Liao, Jing Xiao, Zheng Wang, Chia-Wen Lin, and Shin'ichi Satoh. Uncertainty-aware semantic guidance and estimation for image inpainting. *IEEE Journal of Selected Topics in Signal Processing*, 15(2):310–323, 2021.

Anji Liu, Mathias Niepert, and Guy Van den Broeck. Image inpainting via tractable steering of diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024a.

Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 85–100, 2018.

Haipeng Liu, Yang Wang, Biao Qian, Meng Wang, and Yong Rui. Structure matters: Tackling the semantic discrepancy in diffusion models for image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8038–8047, 2024b.

Chen Long, Wenxiao Zhang, Zhe Chen, Haiping Wang, Yuan Liu, Zhen Cao, Zhen Dong, and Bisheng Yang. SparseDC: Depth completion from sparse and non-uniform inputs. *Information Fusion*, 2024.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11461–11471, 2022.

Xin Ma, Xiaoqiang Zhou, Huaibo Huang, Gengyun Jia, Yaohui Wang, Xinyuan Chen, and Cunjian Chen. Uncertainty-aware image inpainting with adaptive feedback network. *Expert Systems with Applications*, 235:121148, 2024.

Hayk Manukyan, Andranik Sargsyan, Barsegh Atanyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. HD-Painter: High-resolution and prompt-faithful text-guided image inpainting with diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for GANs do actually converge? In *International Conference on Machine Learning (ICML)*, pp. 3481–3490, 2018.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. EdgeConnect: Structure guided image inpainting using edge prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2019.

Mengye Ren, Andrei Pokrovsky, Bin Yang, and Raquel Urtasun. SBNet: Sparse blocks network for fast inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8711–8720, 2018.

Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. In *International Conference on Learning Representations (ICLR)*, 2022.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with Fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2149–2159, 2022.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8798–8807, 2018.

Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4471–4480, 2019.

Yu Zeng, Zhe Lin, Jimei Yang, Jianming Zhang, Eli Shechtman, and Huchuan Lu. High-resolution image inpainting with iterative confidence feedback and guided upsampling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

Lei Zhao, Qihang Mo, Sihuan Lin, Zhizhong Wang, Zhiwen Zuo, Haibo Chen, Wei Xing, and Dongming Lu. UCTGAN: Diverse image inpainting based on unsupervised cross-space translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2017.

A Complete quantitative results

Tables 4–5 give every method of Table 1 on all five mask buckets; figure 5 draws the same numbers. Table 6 is the 200-image diffusion comparison and Table 8 the per-bucket depth statistics of the learned policy.

Table 4: All five mask buckets for every method in Table 1 (PSNR↑ and SSIM↑): 1,500 test images per bucket, identical fixed free-form masks and metric code. Baselines use their official pretrained weights; UAI-Net rows use learned halting; TTA and the Places2 "+ more data" fine-tune are as in Table 1. Bold: best per dataset and bucket, TTA excluded.

		10–20%		20–30%		30–40%		40–50%		50–60%	
Dataset	Method	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
CelebA-HQ	DeepFillv2	30.78	0.953	27.72	0.919	25.73	0.882	24.16	0.843	22.80	0.799
	LaMa	30.72	0.954	27.52	0.918	25.49	0.881	23.87	0.840	22.45	0.795
	MISF	31.54	0.957	28.32	0.924	26.27	0.889	24.63	0.850	23.19	0.807
	Latent codes	31.75	0.956	28.90	0.924	27.00	0.891	25.39	0.855	23.79	0.814
	HINT	32.50	0.962	29.28	0.932	27.20	0.901	25.58	0.867	24.14	0.829
	SPA-Former	**32.85**	**0.964**	**29.70**	**0.937**	**27.65**	**0.907**	**25.98**	**0.875**	**24.48**	**0.838**
	UAI-Net, PSNR-opt.	31.52	0.959	28.22	0.927	26.13	0.892	24.44	0.854	22.96	0.811
	UAI-Net	31.55	0.958	28.26	0.925	26.12	0.888	24.41	0.849	22.90	0.804
	UAI-Net + TTA	31.86	0.961	28.57	0.929	26.46	0.895	24.78	0.858	23.28	0.816
Places2	DeepFillv2	27.06	0.928	24.13	0.875	22.14	0.819	20.63	0.759	19.34	0.695
	LaMa	27.86	0.932	24.99	0.883	23.06	0.831	21.60	0.777	20.35	0.718
	MISF	28.02	0.932	25.10	0.882	23.11	0.829	21.58	0.773	20.27	0.712
	Latent codes	26.03	0.920	23.04	0.863	21.00	0.802	19.44	0.738	18.05	0.669
	HINT	**29.01**	**0.941**	**25.99**	**0.896**	**23.96**	**0.848**	**22.41**	**0.796**	**21.06**	**0.740**
	SPA-Former	28.74	0.939	25.81	0.893	23.78	0.844	22.23	0.792	20.90	0.735
	UAI-Net	27.73	0.931	24.57	0.879	22.37	0.821	20.70	0.762	19.30	0.698
	UAI-Net + more data	27.96	0.934	24.76	0.882	22.52	0.826	20.81	0.767	19.34	0.704
	UAI-Net + TTA	28.05	0.934	24.96	0.884	22.82	0.829	21.19	0.771	19.82	0.709
FFHQ	DeepFillv2†	29.74	0.949	26.72	0.911	24.72	0.870	23.20	0.827	21.84	0.779
	LaMa	29.86	0.950	26.72	0.913	24.64	0.872	23.08	0.830	21.73	0.783
	MISF†	30.35	0.952	27.20	0.916	25.13	0.876	23.56	0.834	22.13	0.787
	HINT†	**31.37**	**0.959**	**28.11**	**0.926**	**25.97**	**0.891**	**24.38**	0.854	**22.94**	**0.813**
	SPA-Former†	31.24	0.958	28.10	0.926	25.96	0.891	24.36	**0.854**	22.91	0.813
	UAI-Net	30.66	0.956	27.33	0.920	25.12	0.881	23.44	0.838	21.89	0.790
	UAI-Net + TTA	30.96	0.958	27.65	0.924	25.47	0.887	23.83	0.848	22.31	0.802

Table 5: All five mask buckets for every method in Table 1 (LPIPS↓ and FID↓): 1,500 test images per bucket, identical fixed free-form masks and metric code. Baselines use their official pretrained weights; UAI-Net rows use learned halting; TTA and the Places2 "+ more data" fine-tune are as in Table 1. Bold: best per dataset and bucket, TTA excluded.

		10–20%		20–30%		30–40%		40–50%		50–60%	
Dataset	Method	LPIPS	FID	LPIPS	FID	LPIPS	FID	LPIPS	FID	LPIPS	FID
CelebA-HQ	DeepFillv2	0.0314	6.03	0.0554	10.32	0.0805	15.41	0.1083	21.76	0.1399	29.28
	LaMa	0.0216	4.18	0.0392	6.78	0.0589	9.59	0.0814	12.54	0.1081	16.02
	MISF	0.0200	3.86	0.0357	6.19	0.0527	8.53	0.0722	10.84	0.0949	13.60
	Latent codes	0.0191	3.83	0.0327	6.07	0.0473	8.08	0.0638	10.16	0.0839	12.51
	HINT	0.0158	3.23	0.0284	5.34	0.0421	7.48	0.0573	9.66	**0.0751**	12.08
	SPA-Former	**0.0153**	**3.04**	**0.0274**	**4.87**	**0.0409**	**6.75**	**0.0565**	**8.69**	0.0753	**10.73**
	UAI-Net, PSNR-opt.	0.0247	4.78	0.0429	7.76	0.0624	10.88	0.0844	14.23	0.1108	18.52
	UAI-Net	0.0211	4.14	0.0380	6.95	0.0563	9.44	0.0771	12.18	0.1020	15.34
	UAI-Net + TTA	0.0219	4.16	0.0385	6.89	0.0562	9.29	0.0761	12.00	0.0998	15.19
Places2	DeepFillv2	0.0481	9.46	0.0831	15.19	0.1218	21.57	0.1617	28.41	0.2064	36.71
	LaMa	0.0350	6.60	0.0617	10.64	0.0908	14.73	0.1218	18.70	0.1569	22.71
	MISF	0.0388	7.47	0.0678	12.02	0.0996	16.79	0.1332	21.46	0.1714	27.08
	Latent codes	0.0457	8.24	0.0807	13.25	0.1189	17.97	0.1602	22.94	0.2074	27.92
	HINT	**0.0307**	**5.61**	**0.0545**	**9.21**	**0.0808**	**12.91**	**0.1088**	**16.57**	**0.1412**	**20.32**
	SPA-Former	0.0318	5.71	0.0566	9.35	0.0844	13.09	0.1140	16.88	0.1490	20.98
	UAI-Net	0.0423	8.16	0.0735	13.11	0.1093	18.58	0.1480	24.06	0.1909	30.40
	UAI-Net + more data	0.0398	7.59	0.0691	12.21	0.1033	17.23	0.1403	22.21	0.1816	27.70
	UAI-Net + TTA	0.0444	8.38	0.0746	13.41	0.1080	18.79	0.1440	24.18	0.1848	30.55
FFHQ	DeepFillv2†	0.0346	8.95	0.0621	14.83	0.0914	21.92	0.1228	30.35	0.1593	40.88
	LaMa	0.0233	5.78	0.0433	9.37	0.0652	12.99	0.0894	16.89	0.1184	21.44
	MISF†	0.0237	6.03	0.0428	9.87	0.0636	14.00	0.0864	18.50	0.1133	24.27
	HINT†	**0.0175**	**4.48**	**0.0324**	**7.28**	**0.0486**	**9.87**	**0.0659**	**12.30**	**0.0866**	**15.13**
	SPA-Former†	0.0194	4.99	0.0352	8.18	0.0528	11.42	0.0721	14.84	0.0955	18.99
	UAI-Net	0.0223	5.82	0.0409	9.39	0.0616	12.90	0.0843	16.61	0.1117	21.07
	UAI-Net + TTA	0.0231	5.78	0.0415	9.32	0.0614	12.69	0.0829	16.26	0.1089	20.45

Table 6: StrDiffusion (Liu et al., 2024b) versus UAI-Net on Places2, both evaluated on the *same* 200 test images and fixed masks per bucket (StrDiffusion's sampler needs $\sim$400 denoising steps, 72 s per image on our GPU, so the full 1,500-image protocol was not feasible). StrDiffusion uses the authors' released Places2 weights (T=400); no CelebA-HQ or FFHQ model was released, so it is compared on Places2 only. FID on 200 images is strongly biased upward and is comparable only within this table, not with Tables 1 or 5. Bold: better of the two.

	StrDiffusion				UAI-Net			
Bucket	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
10–20%	26.73	0.927	**0.0427**	**21.4**	**27.31**	**0.930**	0.0430	22.1
20–30%	23.82	0.874	0.0767	36.4	**24.31**	**0.877**	**0.0750**	**35.0**
30–40%	21.83	0.819	0.1140	52.0	**22.14**	**0.820**	**0.1104**	**50.7**
40–50%	20.15	0.757	0.1572	70.3	**20.46**	**0.759**	**0.1504**	**66.9**
50–60%	18.84	0.690	0.2037	87.1	**19.10**	**0.694**	**0.1937**	**83.7**

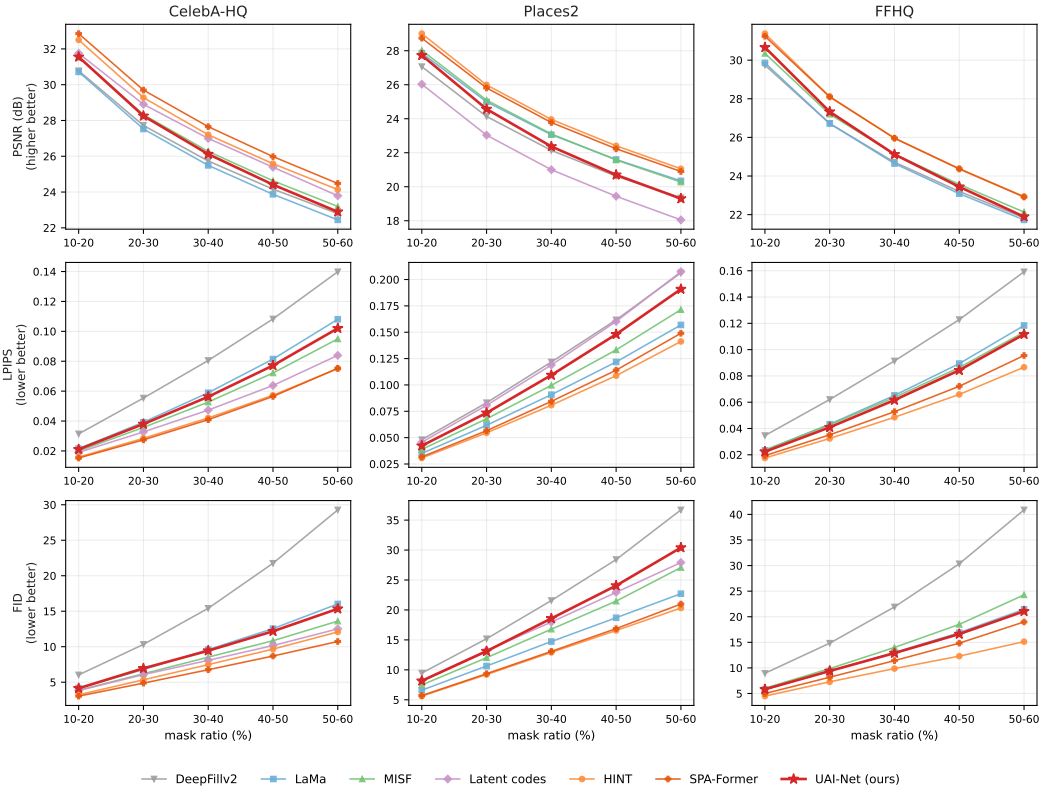


Figure 5: **Fidelity against mask ratio**, every method of Table 1 on the same 1,500 test images and fixed masks per bucket. This is Tables 4–5 drawn rather than tabulated. UAI-Net tracks MISF closely on all three datasets and is clearly ahead of DeepFillv2 and LaMa on the face datasets, while HINT and SPA-Former stay above it throughout; on Places2 it leads the latent-code model in PSNR by a growing margin but trails it in LPIPS at the largest masks. We show the gap rather than crop it: the claim of this paper is the uncertainty behaviour of Figures 11 and 12, not fidelity.

B ABLATIONS AND STATISTICAL TESTS

Table 8 reports the learned policy bucket by bucket. The ablations of Table 3 (uniform, scheduled, learned, $K{=}1$, no-U) differ by small margins, so Table 10 tests them pairwise on the identical 1,500 images and fixed masks: per-image PSNR and LPIPS were saved and compared with a paired Wilcoxon signed-rank test (p_W). Positive ΔPSNR and negative ΔLPIPS favour the first model. No correction for multiple comparisons is applied; under a Bonferroni threshold of $0.05/30 = 1.7{\times}10^{-3}$ over the 30 tests of Table 10, 22 remain significant (9 of 10 for full vs no-U, 8 of 10 for full vs $K{=}1$, 5 of 10 for $K{=}1$ vs no-U).

Table 7: Depth policies on all five mask buckets (CelebA-HQ, 1,500 images per bucket, one set of trained weights; only the test-time halting rule changes). Learned halting is better than uniform $K{=}4$ in every one of the fifteen bucket–metric cells, and better than the scheduled rule $\mathrm{round}(1{+}3U)$ in PSNR and LPIPS in all five buckets while losing to it in FID in three. Table 3 is the 10–20% and 50–60% rows of this table.

	Uniform, all 4 steps			Scheduled $\mathrm{round}(1{+}3U)$			**Learned halting**		
Bucket	PSNR↑	LPIPS↓	FID↓	PSNR↑	LPIPS↓	FID↓	PSNR↑	LPIPS↓	FID↓
10–20%	31.50	0.02153	4.32	31.54	0.02108	4.15	**31.55**	**0.02106**	**4.14**
20–30%	28.21	0.03866	7.25	28.24	0.03803	6.96	**28.26**	**0.03798**	**6.95**
30–40%	26.07	0.05721	9.79	26.11	0.05633	**9.42**	**26.12**	**0.05625**	9.44
40–50%	24.37	0.07829	12.64	24.40	0.07727	**12.15**	**24.41**	**0.07715**	12.18
50–60%	22.84	0.10335	15.94	22.88	0.10217	**15.27**	**22.90**	**0.10201**	15.34
Mean steps		4.00			2.70			**2.54**	

Table 8: Learned depth statistics per bucket (1,500 images): mean steps per hole pixel, tercile means, and within-image Spearman ρ between U and error.

Bucket	Steps	lo / mid / hi tercile	ρ_{img}
10–20%	2.60	1.73 / 2.81 / 3.26	0.669
20–30%	2.56	1.70 / 2.77 / 3.20	0.674
30–40%	2.53	1.69 / 2.74 / 3.16	0.672
40–50%	2.51	1.68 / 2.73 / 3.11	0.664
50–60%	2.48	1.67 / 2.71 / 3.07	0.651

Table 9: AUROC with 95% bootstrap confidence intervals over images (100 resamples, 200 hole pixels per image), for every uncertainty source and mask bucket. These are the intervals behind Table 2, which reports the five-bucket averages. In every dataset and every bucket the interval of the UAI-Net variance head lies entirely above the interval of every other source.

Dataset	Uncertainty source	10-20%	20-30%	30-40%	40-50%	50-60%
CelebA-HQ	UAI-Net variance head σ^2	0.885 [0.880,0.889]	0.878 [0.870,0.879]	0.875 [0.872,0.880]	0.871 [0.866,0.873]	0.868 [0.865,0.872]
	UAI-Net routing map U	0.829 [0.825,0.833]	0.820 [0.814,0.823]	0.813 [0.810,0.820]	0.806 [0.800,0.808]	0.799 [0.796,0.804]
	UAI-Net flip disagreement	0.813 [0.807,0.816]	0.805 [0.798,0.808]	0.803 [0.799,0.810]	0.802 [0.794,0.805]	0.798 [0.793,0.801]
	UAI-Net snapshot ensemble	0.832 [0.827,0.836]	0.826 [0.818,0.828]	0.823 [0.820,0.829]	0.820 [0.815,0.823]	0.819 [0.815,0.824]
	LaMa flip disagreement	0.779 [0.771,0.785]	0.774 [0.767,0.780]	0.767 [0.764,0.775]	0.761 [0.753,0.764]	0.754 [0.747,0.758]
	HINT flip disagreement	0.808 [0.802,0.813]	0.801 [0.797,0.807]	0.796 [0.792,0.804]	0.795 [0.787,0.797]	0.790 [0.787,0.797]
	SPA-Former flip disagreement	0.788 [0.781,0.793]	0.786 [0.780,0.790]	0.784 [0.779,0.789]	0.781 [0.772,0.782]	0.779 [0.772,0.781]
	distance to known pixels	0.669 [0.662,0.677]	0.667 [0.661,0.675]	0.669 [0.663,0.678]	0.673 [0.666,0.678]	0.674 [0.670,0.681]
Places2	UAI-Net variance head σ^2	0.863 [0.855,0.871]	0.858 [0.852,0.865]	0.856 [0.849,0.862]	0.856 [0.850,0.862]	0.855 [0.850,0.863]
	UAI-Net routing map U	0.766 [0.758,0.775]	0.766 [0.760,0.777]	0.765 [0.758,0.773]	0.760 [0.754,0.766]	0.758 [0.754,0.767]
	UAI-Net flip disagreement	0.795 [0.784,0.799]	0.792 [0.786,0.802]	0.792 [0.781,0.799]	0.790 [0.782,0.798]	0.784 [0.777,0.793]
	UAI-Net snapshot ensemble	0.812 [0.805,0.822]	0.806 [0.801,0.816]	0.803 [0.793,0.810]	0.800 [0.791,0.806]	0.795 [0.787,0.804]
	LaMa flip disagreement	0.788 [0.778,0.794]	0.778 [0.770,0.787]	0.772 [0.764,0.782]	0.767 [0.758,0.774]	0.760 [0.752,0.767]
	HINT flip disagreement	0.814 [0.807,0.823]	0.808 [0.799,0.814]	0.803 [0.797,0.814]	0.799 [0.791,0.808]	0.792 [0.784,0.798]
	SPA-Former flip disagreement	0.794 [0.785,0.802]	0.790 [0.784,0.801]	0.785 [0.776,0.792]	0.778 [0.768,0.781]	0.772 [0.764,0.780]
	distance to known pixels	0.620 [0.613,0.629]	0.615 [0.609,0.623]	0.617 [0.614,0.627]	0.619 [0.610,0.626]	0.622 [0.615,0.626]
FFHQ	UAI-Net variance head σ^2	0.880 [0.874,0.885]	0.876 [0.871,0.880]	0.874 [0.869,0.877]	0.871 [0.866,0.874]	0.868 [0.865,0.873]
	UAI-Net routing map U	0.818 [0.814,0.825]	0.812 [0.806,0.818]	0.808 [0.804,0.816]	0.801 [0.794,0.804]	0.792 [0.788,0.797]
	UAI-Net flip disagreement	0.813 [0.807,0.819]	0.811 [0.806,0.818]	0.811 [0.806,0.816]	0.808 [0.801,0.810]	0.804 [0.799,0.809]
	UAI-Net snapshot ensemble	0.841 [0.835,0.844]	0.837 [0.833,0.842]	0.836 [0.829,0.838]	0.832 [0.825,0.834]	0.828 [0.826,0.836]
	LaMa flip disagreement	0.774 [0.765,0.777]	0.765 [0.760,0.773]	0.761 [0.756,0.768]	0.755 [0.748,0.759]	0.750 [0.746,0.756]
	distance to known pixels	0.686 [0.678,0.693]	0.685 [0.682,0.694]	0.685 [0.679,0.692]	0.682 [0.672,0.686]	0.687 [0.679,0.695]

Table 10: Paired tests on CelebA-HQ, n=1,500 per bucket.

Comparison	Bucket	ΔPSNR	p_W	ΔLPIPS	p_W
Full vs no-U	10–20%	$+0.065$	3.8×10^{-6}	-0.0002	2.0×10^{-3}
	20–30%	$+0.089$	5.2×10^{-14}	-0.0003	4.0×10^{-4}
	30–40%	$+0.087$	4.7×10^{-13}	-0.0008	3.4×10^{-12}
	40–50%	$+0.092$	2.2×10^{-14}	-0.0011	2.6×10^{-18}
	50–60%	$+0.073$	7.1×10^{-11}	-0.0014	1.8×10^{-16}
Full vs K=1	10–20%	$+0.022$	0.080	-0.0001	1.2×10^{-3}
	20–30%	$+0.047$	2.8×10^{-4}	-0.0002	5.7×10^{-3}
	30–40%	$+0.064$	9.6×10^{-8}	-0.0004	8.6×10^{-5}
	40–50%	$+0.067$	7.4×10^{-9}	-0.0005	3.5×10^{-4}
	50–60%	$+0.064$	6.1×10^{-9}	-0.0007	7.3×10^{-5}
K=1 vs no-U	10–20%	$+0.044$	4.5×10^{-4}	-0.0001	0.38
	20–30%	$+0.042$	2.4×10^{-4}	-0.0001	0.18
	30–40%	$+0.023$	0.033	-0.0004	4.0×10^{-4}
	40–50%	$+0.025$	0.032	-0.0006	2.6×10^{-5}
	50–60%	$+0.008$	0.31	-0.0007	5.2×10^{-5}

C QUALITATIVE COMPARISONS

Figure 3 in the main text gives a compact version of these grids. Below, every method is shown on every dataset at three mask ratios, followed by the images where UAI-Net does worst.

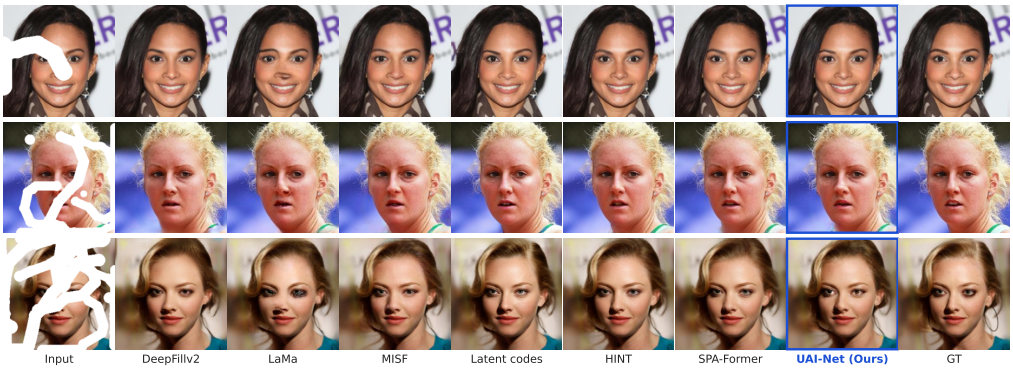


Figure 6: **Qualitative comparison on CelebA-HQ.** Every method on the shared fixed masks, one row per mask bucket (10–20%, 30–40%, 50–60%); hole in white, our column boxed. Each row is the image at the median of UAI-Net's per-image PSNR over the 1,500 images of that bucket, not a selected example, and each panel is the stored output that produced Table 1.

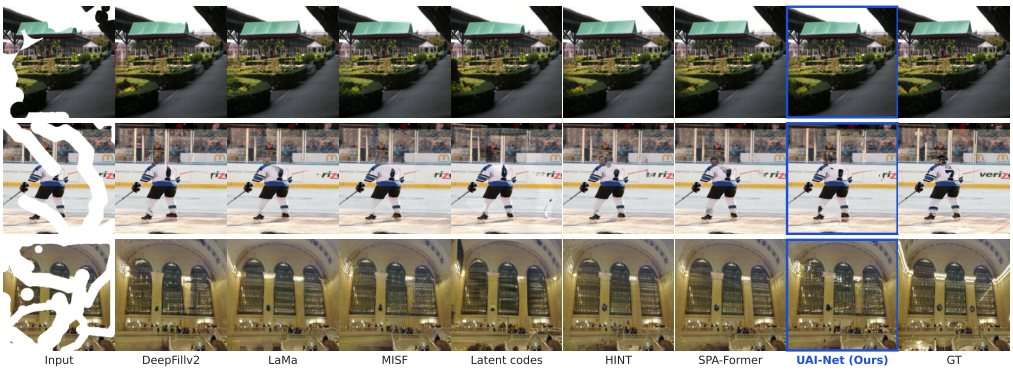


Figure 7: **Qualitative comparison on Places2.** Same protocol as Figure 6. Places2 is where UAI-Net trails HINT and SPA-Former by the widest margin (Table 1); these rows are what that margin looks like.

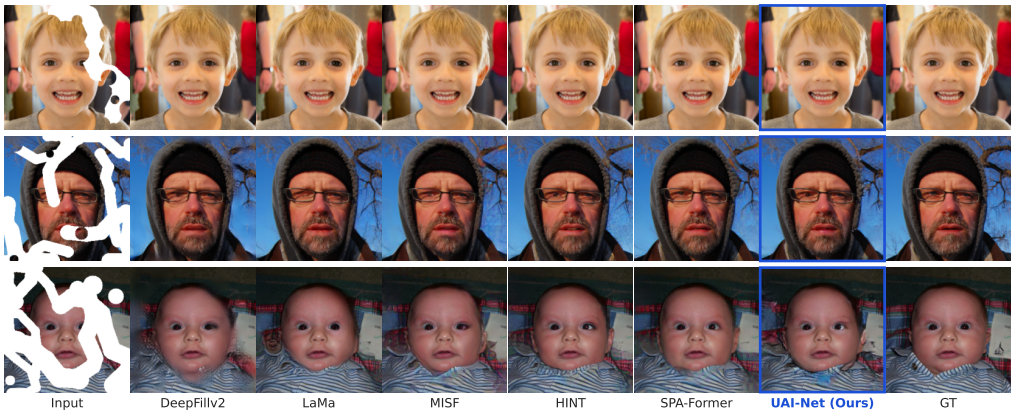


Figure 8: **Qualitative comparison on FFHQ.** Same protocol as Figure 6. No FFHQ weights were released for DeepFillv2, MISF, HINT or SPA-Former, so their CelebA-HQ weights are applied († in Table 1); the latent-code model has no FFHQ release and is therefore absent.

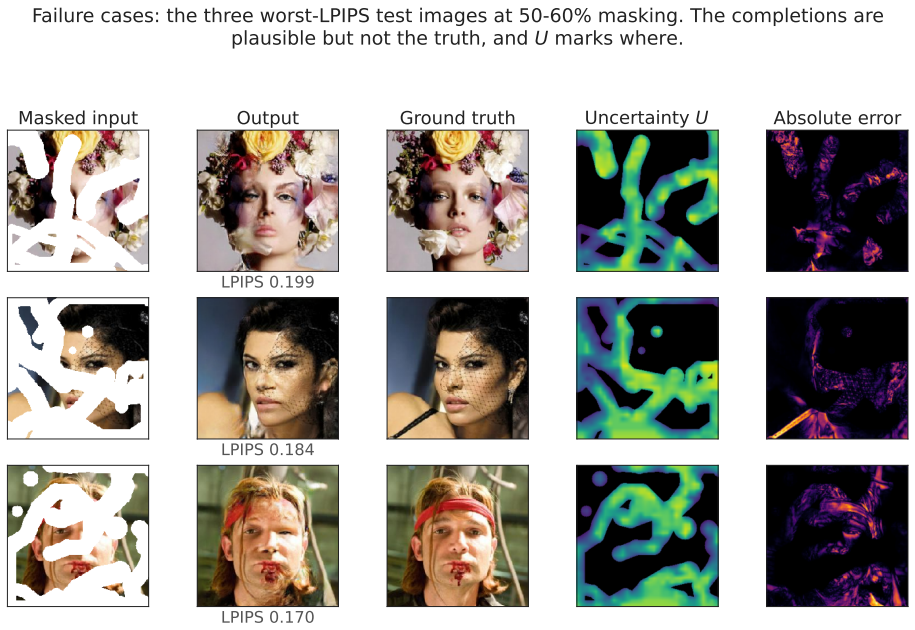


Figure 9: **Failure cases:** the three worst-LPIPS test images at 50–60% masking. The completions are plausible but not the truth, and U marks where: the bright regions of the uncertainty map coincide with the regions the model invented.

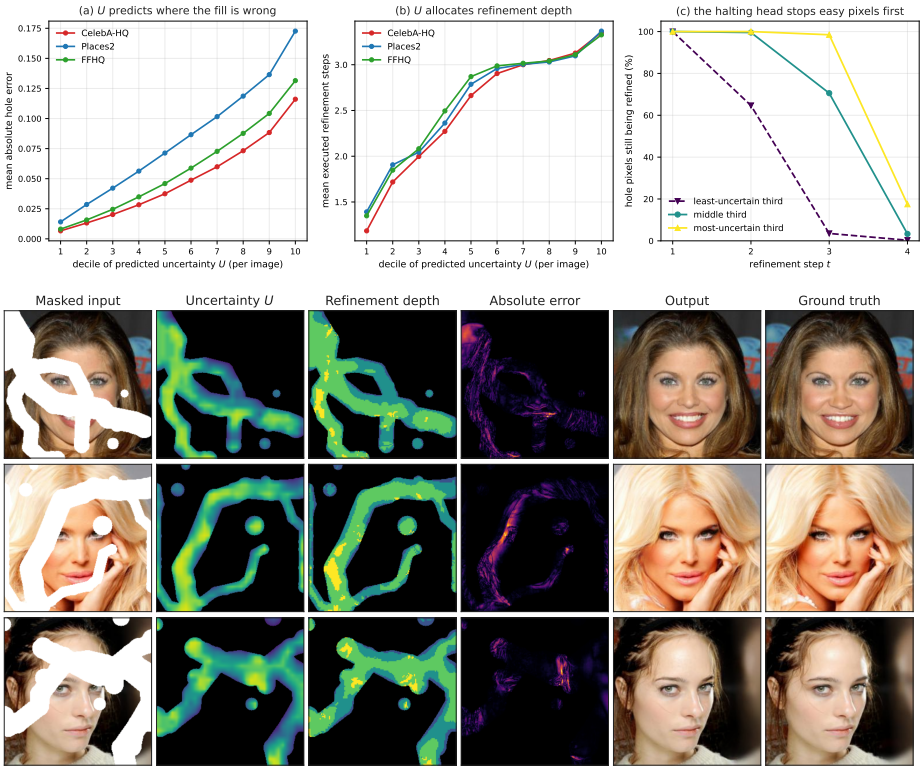


Figure 10: **The two claims of the paper, drawn** (30–40% masks, 300 test images per dataset, hole pixels only; deciles and terciles are taken within each image, so a global offset in U cannot produce the trend). **(a)** Mean absolute error rises monotonically with the decile of U, by more than $12\times$ on every dataset (Places2 $12.3\times$, FFHQ $16.2\times$, CelebA-HQ $17.5\times$) from the least- to the most-uncertain decile: the map is trained only on the rank of the model's own error and recovers that ordering at test time. **(b)** The executed refinement depth rises with the same decile, from 1.2–1.4 steps to $\sim$3.35, on all three datasets: refinement depth follows predicted difficulty. Note that depth is decided by the halting probabilities p_t, not computed from U (figure 2); the monotone relation is learned, not imposed. **(c)** Survival of the refinement loop on CelebA-HQ: after two steps only 3.5% of the least-uncertain third is still being refined, against 70.6% of the middle third and 98.5% of the most-uncertain third. This is the mechanism behind the 2.54-step average of Table 3. The three-row panel below is the full version of the qualitative row of Figure 4.

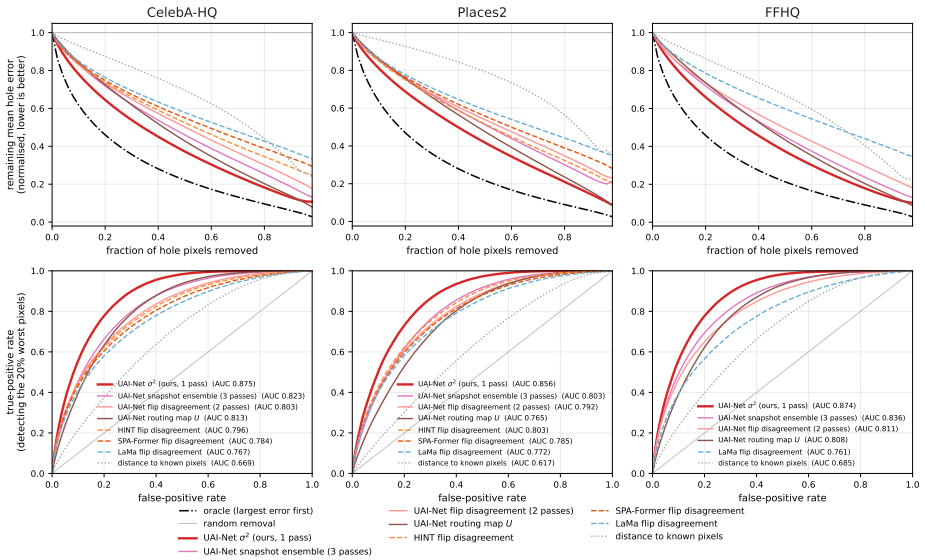


Figure 11: **Sparsification and error-detection curves** (30–40% masks, 500 test images, hole pixels only). **Top:** mean hole error remaining after removing the most-uncertain fraction of hole pixels, normalised to 1 at $f{=}0$ (Ilg et al., 2018); the oracle removes the largest-error pixels first and the flat line is random removal. A curve closer to the oracle is a better uncertainty estimate. **Bottom:** ROC for detecting the 20% worst-error hole pixels of each method's own output, with the area under each curve in the legend. The variance head of UAI-Net is closest to the oracle and highest on the ROC on all three datasets, using one forward pass against the two of flip disagreement and the three of the snapshot ensemble. The AUCs shown here are from the 30–40% mask bucket; Table 2 reports the corresponding five-bucket averages and Table 9 the per-bucket intervals.

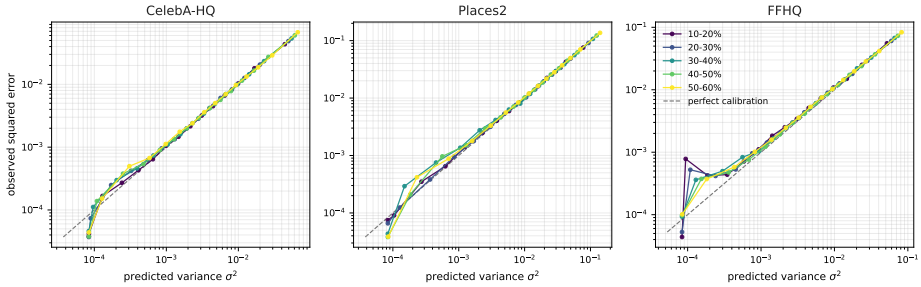


Figure 12: **Reliability of the predicted variance.** Hole pixels are sorted by predicted variance $\sigma^2 = e^{\ell}$ and split into 15 equal-count bins; each point is one bin, plotted as mean predicted variance against mean observed squared error (log–log, $[0, 1]$ intensity units). A perfectly calibrated head lies on the diagonal. The head tracks the diagonal across three orders of magnitude on all three datasets and all five mask ratios, over-estimating slightly only in the lowest-variance bins, where the absolute error is $\sim 10^{-4}$. This is what the UCE values of §4.3 measure. The compared baseline uncertainty measures are disagreement statistics rather than a predictive variance, so this calibration curve is not defined for them.

D ADAPTIVE REFINEMENT BEHAVIOUR

All panels below are rendered by `make_appendix_grids.py` from the trained CelebA-HQ checkpoint; no image is edited by hand.

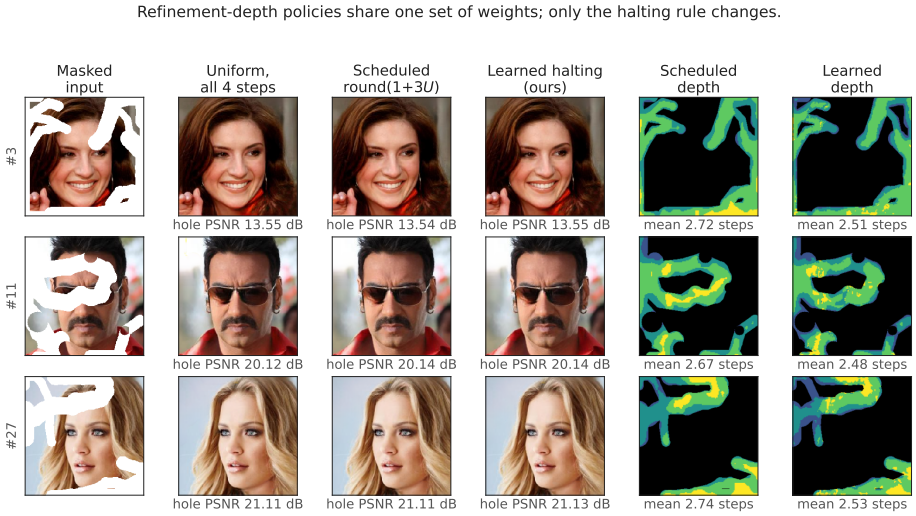


Figure 13: **The three depth policies share one set of weights**; only the halting rule changes at test time. Outputs are visually indistinguishable and hole-PSNR differs by at most 0.02 dB, while the learned policy uses 0.2 fewer steps per hole pixel than the scheduled one. Depth maps: viridis, 1 (dark) to $K{=}4$ (yellow); known pixels are black. Masks are shown in white.

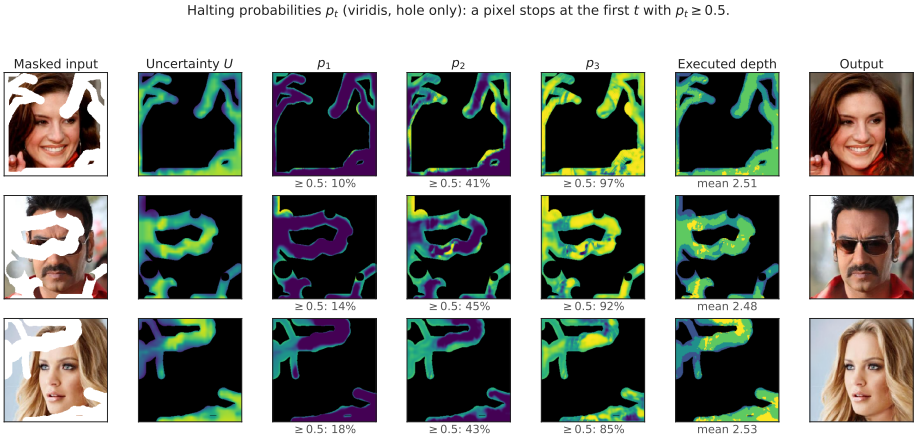


Figure 14: **Halting probabilities per refinement step.** p_t rises monotonically with t for most pixels, so the fraction of hole pixels that have reached $p_t \geq 0.5$ grows from $\sim$14% at $t{=}1$ to $\sim$90% at $t{=}3$; the pixels that survive to step 4 are the high-U ones (hair, eyes, mouths). The head is evaluated at the refinement resolution $128{\times}128$ and shown upsampled.

E HALTING-LOGIT COLLAPSE

Figure 15 is the evidence behind §3.5: the logit distribution of the halting head under the continuous depth target and under the integer target with confidence penalty, measured on the same 50 test images.

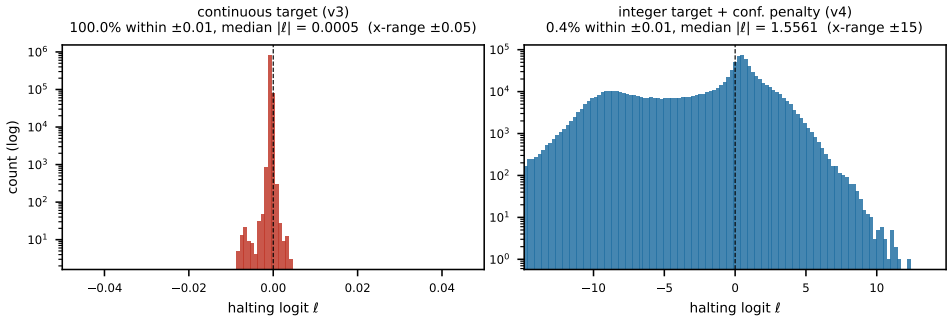


Figure 15: **Halting-logit collapse.** Halting-head logits over all hole cells and halting steps of 50 CelebA-HQ test images (30–40% masks, 0.9M logits per model, log-scale counts). Left: continuous depth target, final checkpoint (60 epochs); 100.0% of logits lie within ± 0.01 (note the ± 0.05 x-range). Right: integer target with confidence penalty (UAI-Net); 0.4% within ± 0.01, median $|\ell| = 1.56$.

F Implementation and reproducibility

Commands. Training: `python uai_net.py --dataset celeba_hq --mode train --tag v4 --base_ch 96 --epochs 50`. Evaluation of UAI-Net as reported everywhere in the paper: `--mode eval --eval_n 1500 --halt_mode learned` (`learned` is the default, so the flag may be omitted); add `--tta` for the TTA row, and substitute `--halt_mode usched` for the scheduled row or `--halt_mode full` for the uniform four-step row of Table 3. Ablations: `--ponder_K 1`; `--routing none --lam_unc 0 --lam_ponder 0`. Places2 "+ more data": `--dataset places2 --mode train --tag v4ft --base_ch 96 --epochs 6 --train_size 300000 --finetune <v4 best.pth>`, which loads the Places2 weights and starts a fresh learning-rate schedule. *UAI-Net variants of Table 1.* "PSNR-opt." uses λ_{rec}=6, λ_{adv}=0.2, a PSNR-selected checkpoint and c=64; "+ more data" fine-tunes the Places2 model for six further epochs on 300,000 training images; flip TTA averages the prediction with its mirrored counterpart. *Loss terms.* $\mathcal{L}_1$ is a masked L1 on the composite plus $0.5\times$ the same loss on rgb_0; $\mathcal{L}_{\text{perc}}$ and $\mathcal{L}_{\text{sty}}$ are feature and Gram-matrix losses (Johnson et al., 2016; Gatys et al., 2016) on a dilated ImageNet ResNet-50, i.e. the high-receptive-field perceptual loss of Suvorov et al. (2022); $\mathcal{L}_{\text{adv}}$ is a hinge loss against a spectral-normalised PatchGAN (Miyato et al., 2018) with R_1 regularisation (Mescheder et al., 2018), and $\mathcal{L}_{\text{fm}}$ matches its intermediate activations (Wang et al., 2018); the last two are switched on after three epochs. *Heteroscedastic loss.* The NLL is back-propagated into the prediction, where its gradient is scaled by $e^{-\ell}$ and is weakest on the high-variance pixels that most need correction (Seitzer et al., 2022); restricting it to the variance head is untested. *FFHQ run.* Training was interrupted once after 15 epochs and resumed from the saved weights with the learning-rate schedule restored. *Masks.* Free-form masks are generated once per (image index, bucket) with a fixed seed and cached; all methods and variants are evaluated on the identical set. *Baselines.* `compare_uai.py` exports the 1,500 test images and masks of each bucket, runs each baseline's official inference code with its released weights, composites the known pixels and scores the outputs with the same metric functions as UAI-Net. *Figures.* All figures showing model outputs are rendered by scripts from the trained checkpoints: Figure 2 by `make_fig2_mechanism.py`, which replays the hard-halting refinement loop and checks that it reproduces the model output (maximum absolute difference 1.2×10^{-7}) before drawing; Figure 15 by `make_fig4_collapse.py`; and Figure 4 by `make_fig3_uncertainty.py`. *Hyper-parameters.* Learning rate 2×10^{-4} (generator, AdamW, $\beta = (0.5, 0.999)$) and 10^{-4} (discriminator, $\beta = (0, 0.99)$), batch 8, cosine decay to 10^{-6}, EMA 0.999, gradient clipping 1.0, R_1 weight 10 every 16 discriminator steps, adversarial and feature-matching losses enabled after the first three epochs and ramped up linearly over the next three.

G POSITIONING AGAINST CONFIDENCE-FEEDBACK AND ADAPTIVE-COMPUTATION METHODS

Table 11 summarises, as described in the respective papers, how UAI-Net differs from inpainting methods that feed a confidence or uncertainty map back into the network and from adaptive-computation methods. It records design choices, not quality; a dash means the element is not part of the method.

Table 11: Design comparison. "Map": the per-pixel quantity the method predicts; "passes": how many times the hole is processed; "depth": whether the number of passes varies per pixel and is trained; "routing": whether the map selects between feature sources; "evaluated": whether the map is scored against the true reconstruction error.

Method	Map	Passes	Per-pixel depth	Routing	Evaluated
Zeng et al. (2020)	confidence	4, fixed	heuristic (re-generate if confidence rose)	–	–
U^2AFN (Ma et al., 2024)	aleatoric variance	4, fixed	– (high variance re-marked missing)	–	–
SeGuE-Net (Liao et al., 2021)	segmentation + uncertainty	fixed	–	–	–
PSM (Li et al., 2024)	mean + variance	fixed	– (low-variance pixels kept)	attention weights	–
ACT / SACT (Graves, 2016; Figurnov et al., 2017)	–	learned	SACT: per position (ponder cost)	–	n/a
PonderNet (Banino et al., 2021)	–	learned	per sample	–	n/a
UAI-Net (ours)	rank of own error (U), σ^2	1–4, learned	integer target from U, hard forward	local / global / boundary	Table 2

Table 12: Training configuration of every UAI-Net model in the paper (identical unless stated).

Resolution / mask	256×256; free-form masks (Yu et al., 2019), hole ratio 10–60%, fixed test masks per (image, bucket)
Architecture	gated-conv U-Net, c=96, 32.4M parameters; U at 32×32; refinement at 128×128, K=4, shared block
Optimiser	AdamW, lr 2×10^{-4} (G, β=$(0.5, 0.999)$, wd 10^{-5}), 10^{-4} (D), cosine decay to 10^{-6}, grad-clip 1.0
Schedule	batch 8; 50 epochs CelebA-HQ (28k images), 30 epochs Places2 / FFHQ (50k images); EMA 0.999; mixed precision
Losses	weights $3, 1, 2, 0.05, 0.02, 2, 120, 0.75, 10$ (§3.6); adversarial and feature-matching terms from epoch 3, ramped over three epochs; R_1 weight 10 every 16 steps
Halting head	integer target $\mathrm{round}(1+(K-1)U)$, confidence penalty, threshold $p_t \geq 0.5$, straight-through backward
Selection / seed	checkpoint with the lowest mean LPIPS of the per-epoch evaluation (48 test images per bucket); seed 1234; one RTX 4090 ($\sim$12 h CelebA-HQ)
Baselines	official released weights, same images, masks and metric code; known pixels composited before scoring