SURGIMIST: CONTROLLED SMOKE SYNTHESIS FOR SURGICAL IMAGE RESTORATION AND PERCEPTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Surgical smoke changes image appearance and obscures cues used by both restoration models and downstream perception systems. Paired clear and smoky observations are difficult to acquire in surgery, while restoration scores alone do not establish whether removing smoke helps a task. We introduce SurgiMist, a controlled synthesis framework that combines relative depth, heterogeneous optical fields, layered accumulation, texture evolution, and scattering-inspired image formation. Its nested configurations support restoration benchmarking, and a geometry-preserving variant transfers existing semantic labels to smoke-augmented images. We evaluate the framework through a fixed restoration benchmark, three restoration backbones on external data, and a video-disjoint surgical segmentation study with three training seeds. On the fixed four-condition restoration benchmark, the complete configuration achieves the best mean PSNR, SSIM, and LPIPS among the four synthesis configurations. For segmentation, its geometry-preserving variant reaches $50.18 \pm 0.12\%$ mean intersection-over-union across 15 balanced synthetic conditions, a 15.71 percentage-point gain over original-image training. Under severe smoke, direct augmentation reaches 50.28%, compared with 39.45% for a learned restoration front end. In a blinded review of 500 paired judgments by five doctors on real-smoke frames, B4 receives slightly more votes than LSD3K for visibility, structure preservation, and viewing preference, and fewer artifact flags. Matched component controls separate the benefit of smoke augmentation from Gaussian smoothing alone. Together, these results position controlled synthesis as a practical interface for measuring restoration quality and training smoke-robust perception, with the downstream objective determining the useful processing route.

1 INTRODUCTION

Surgical smoke is an acquisition disturbance with two consequences: it changes the image that a clinician sees and changes the evidence available to a perception model. These consequences call for different evaluations. A restoration model may produce a visually improved frame without recovering the features needed for segmentation; conversely, a predictor trained with smoke can retain task accuracy without producing a restored image. A useful smoke benchmark should therefore connect image formation, restoration, and downstream learning.

Building this connection from real surgery is difficult. The tissue, camera, instruments, and smoke move together, making pixel-aligned clear and smoky captures scarce. Existing surgical desmoking datasets and recent real-image restoration benchmarks address important parts of this problem (Xia et al., 2024; Chang & Chen, 2024; Pei et al., 2026). Synthetic data offers complementary control: the same source frame can be observed under different optical fields and smoke strengths, and existing annotations can be reused when the transformation preserves geometry. That control is valuable only if the evaluation distinguishes the renderer, the training distribution, and the downstream objective.

We introduce *SurgiMist*, a family of controlled surgical smoke generators and a connected evaluation protocol. The generator combines a depth-dependent transmission model with heterogeneous optical fields, layered accumulation, evolving textures, and scattering-inspired image effects. Four nested configurations, B1–B4, form a restoration benchmark. For semantic learning, a geometry-preserving B4 variant removes coordinate displacement while retaining photometric smoke effects. Mean-

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.

transmittance calibration then places different renderers on a common severity axis, allowing the same held-out frames to be evaluated under matched average attenuation.

The resulting evidence supports two uses of synthesis. First, the complete B4 configuration provides a strong restoration training distribution within the fixed benchmark: its 12k benchmark checkpoint has the best four-condition mean PSNR, SSIM, and LPIPS among B1–B4. Its external behavior depends on the evaluation domain. Across three backbones it lowers NIQE, BRISQUE, and FADE relative to LSD3K training on DeSmoke-LAP, whereas paired external benchmarks often favor LSD3K. Second, smoke augmentation greatly improves segmentation under synthetic degradation. On a balanced 15-condition panel, geometry-preserving B4 improves mean intersection-over-union (mIoU) by 15.71 percentage points over original-only training. The gain is largely shared with simpler smoke augmentations; the contribution is controllable, reusable training variation rather than a claim that every optical component is uniquely necessary.

A task-level comparison makes the practical consequence clear. On severe B4-photo test images, direct augmentation yields 50.28% mIoU, while restoring the image before an original-trained predictor yields 39.45% with the B4-trained restorer. Independent architecture timing also shows the cost of the additional restoration network. Restoration and augmentation therefore serve different objectives, and should be compared through the task they are intended to support.

Our contributions are:

- A controlled smoke synthesis family that exposes depth, spatial heterogeneity, layer structure, texture evolution, and optical image effects, together with a label-preserving variant and severity calibration.
- A restoration study spanning nested synthesis configurations, three backbones, paired and real-image diagnostics, and data-scale controls, with complete condition-level results.
- A video-disjoint, three-seed surgical perception study that measures balanced smoke robustness, resolves component effects, and compares direct augmentation with restoration followed by prediction.

2 Related Work

Surgical smoke data and restoration. Existing resources address complementary aspects of surgical visibility. DeSmoke-LAP learns unpaired translation from real hysterectomy videos using cycle consistency, inter-channel discrepancies, and a dark-channel loss (Pan et al., 2022). Its unpaired setting avoids requiring simultaneous smoky and smoke-free observations, but precludes direct full-reference evaluation. LSD3K provides 3,000 synthetic pairs by combining Cholec80 imagery with smoke masks generated in Blender (Chang & Chen, 2024). DesmokeData instead constructs 961 real-image pairs from relatively stationary prostatectomy sequences and compensates for motion between smoky and reference frames (Xia et al., 2024). More recently, SurgClean extends real surgical restoration to smoke, lens fogging, and splashes across two medical sites, with 3,113 paired-reference images and a benchmark of 22 restoration methods (Pei et al., 2026). These datasets establish the importance of both realistic degradation and explicit reference construction. SurgiMist complements them with controllable smoke synthesis whose components can be varied while retaining the same source image. This enables matched restoration comparisons and perception experiments that distinguish the value of smoke augmentation from that of individual degradation components. Controlled synthetic tests and naturally occurring smoke evaluations consequently address complementary questions.

Controlled degradation and restoration models. The atmospheric scattering formulation relates observed appearance to scene radiance, transmission, and airlight; the dark-channel prior provides a classical inverse method based on outdoor-image statistics (He et al., 2011). HazeFlow combines an ordinary-differential-equation formulation of dehazing with Markov Chain Brownian Motion fields for spatially non-homogeneous haze generation (Shin et al., 2025). This makes spatially varying density a relevant synthesis comparator beyond homogeneous haze. Monocular estimators such as Depth Anything V2 provide depth cues that can condition image formation without requiring instrumented acquisition (Yang et al., 2024). On the restoration side, HINet uses half instance normalization in a multistage network (Chen et al., 2021), Restormer adapts Transformer attention

to high-resolution restoration (Zamir et al., 2022), and NAFNet demonstrates strong restoration with a simplified architecture (Chen et al., 2022). DehazeFormer specifically adapts vision Transformers to dehazing (Song et al., 2023). SurgiMist studies how controllable degradation choices affect these restoration systems and their downstream use. Matched component controls isolate the effects of increased synthesis complexity under a common architecture and training budget.

Surgical perception and evaluation. Cholec80, introduced with EndoNet (Twinanda et al., 2017), and the BernBypass70 and StrasBypass70 components of MultiBypass140 (Lavanchy et al., 2024) provide surgical video sources spanning procedures and centers. CholecSeg8k adds pixel annotations to selected Cholec80 frames (Hong et al., 2020), enabling semantic evaluation with architectures such as DeepLabV3 (Chen et al., 2017). These resources support separate analyses of source diversity and semantic robustness, with pixel-level performance evaluated where the required annotations are available. Image reconstruction is commonly measured through pixel fidelity, structural similarity (Wang et al., 2004), and learned perceptual distances (Zhang et al., 2018). Without paired references, NIQE (Mittal et al., 2013), BRISQUE (Mittal et al., 2012), and FADE (Choi et al., 2015) offer statistical quality or fog-density diagnostics. Their objectives differ from correct anatomical segmentation, and DeSmoke-LAP explicitly cautions against relying solely on reference-free metrics for surgical images (Pan et al., 2022). SurgiMist accordingly evaluates restoration followed by a fixed predictor separately from direct prediction after task-specific augmentation. Together with controlled component comparisons and inference-cost measurements, this separation tests whether a visually improved image also improves the intended perception task.

3 Controlled smoke synthesis

Image formation. Let $J(x) \in [0,1]^3$ be a source RGB image, $d(x) \in [0,1]$ its normalized relative depth, and $\mathcal{F}$ the valid endoscopic field of view. We estimate depth with Depth Anything V2 (Yang et al., 2024); it supplies relative geometry, not metric smoke path length. Given an optical-depth field $\tau(x) \geq 0$, transmission and the attenuated image are

$$T(x) = \exp[-a\tau(x)], \qquad I_0(x) = J(x)T(x) + A[1 - T(x)], \tag{1}$$

where $a > 0$ sets the overall strength and A is the airlight intensity. The source frame is a training reference; an unannotated surgical frame is not assumed to be physically smoke-free.

From heterogeneous fields to layers. A homogeneous control, denoted H, sets $\tau(x) = d(x)$. B1 uses $\tau(x) = \beta(x)\rho(x)d(x)$, with a texture-derived extinction field β and a spatial density field ρ. In the calibrated implementation, normalized texture and density-bank images define these fields; the texture is horizontally flipped and mapped to $\beta = 0.8 + 0.1v$. B2 divides the normalized depth interval into $L = 5$ layers:

$$\tau(x) = \sum_{\ell=1}^{L} \kappa_\ell(x)\, \mathrm{clip}\left(d(x) - \frac{\ell - 1}{L}, 0, \frac{1}{L}\right). \tag{2}$$

Here $\kappa_\ell = \beta_\ell \rho_\ell$. This expression accumulates only the portion of each layer intersected by the relative depth. B3 evolves the texture fields using a simplified advection/diffusion-style update before evaluating Eq. 2. This is a lightweight field construction, not a solved three-dimensional fluid or radiative-transfer system. B4 adds a scattering-inspired Gaussian image operation and a small transmission-dependent coordinate displacement. These configurations expose distinct sources of variation without changing the underlying RGB source.

Geometry-preserving training. Semantic labels should remain aligned with image coordinates. We therefore use *B4-photo*, the B4 branch without coordinate displacement, for the controlled perception study. In its current implementation, scattering is represented by a mask-normalized Gaussian operation

$$I_{\mathrm{photo}} = \frac{G_\sigma(MI_0)}{\max\{G_\sigma(M), \epsilon\}}, \qquad \sigma \sim \mathcal{U}(0.3, 0.9), \tag{3}$$

where M is the field-of-view mask and G_σ uses a 9×9 kernel. Pixels outside the field of view are restored to their original values. The Gaussian is an image-space approximation motivated by

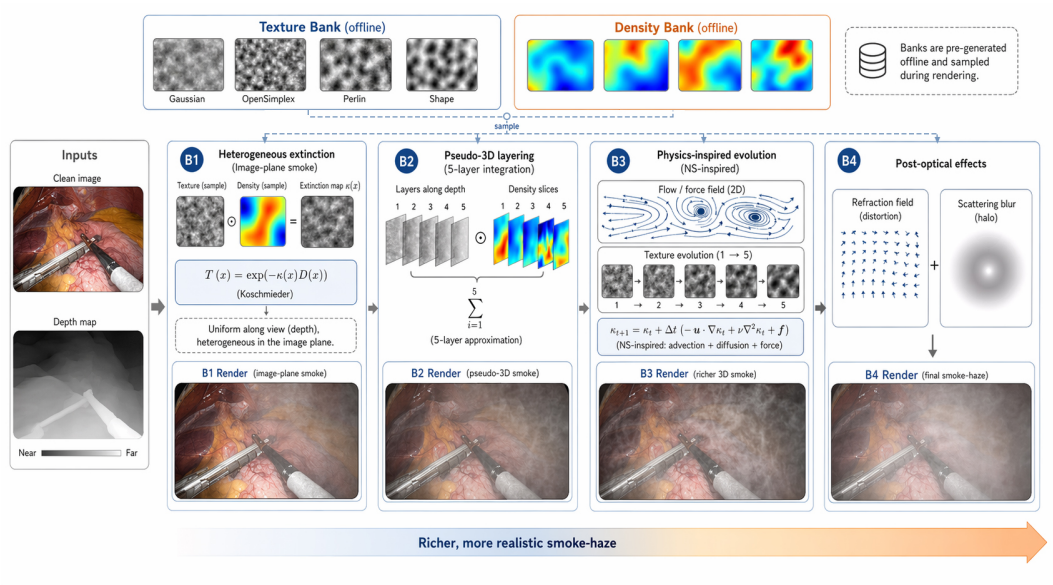


Figure 1: **Staged smoke synthesis in SurgiMist.** Separate texture and density banks drive heterogeneous extinction (B1), five-layer depth accumulation (B2), a simplified advection–diffusion-inspired texture update (B3), and transmission-dependent displacement with global Gaussian softening (B4). The B3 step is procedural rather than a full Navier–Stokes solver. B4-photo, used to retain segmentation-label coordinates, omits displacement.

scattering; its width is global rather than a spatially resolved physical scattering estimate. Full B4 applies a small bilinear displacement depending on $1 - T(x)$ before the Gaussian operation; B4-photo evaluates Eq. 3 without displacement.

Matching severity across renderers. For each image and target mean transmission $\bar{T}$, we choose a by bisection so that

$$\frac{1}{|\mathcal{F}|} \sum_{x \in \mathcal{F}} \exp[-a\tau(x)] = \bar{T}. \tag{4}$$

The current study samples $\bar{T} \sim \mathcal{U}(0.35, 0.85)$ in training and evaluates at 0.85, 0.60, and 0.35; lower values indicate stronger attenuation. Matching the mean fixes a common severity reference while retaining differences in spatial structure. It does not equate every perceptual property of the renderers.

Component controls. We compare B4-photo with H and a matched heterogeneous-field control, MCBM, adapted from the field construction of HazeFlow (Shin et al., 2025). This is not a reproduction of the full HazeFlow training system. To identify which effects support segmentation, we also evaluate H with the same Gaussian operation, B4-photo without that operation, and Gaussian-only augmentation without smoke. The original fixed B1–B4 benchmark and the later severity-calibrated implementation are separate protocols; Appendix A records their parameterization differences.

4 Experimental Design

Restoration benchmark and external evaluation. The fixed restoration benchmark uses Cholec80 (Twinanda et al., 2017) with video-disjoint splits: videos 01–56 supply 5,600 training frames, 57–64 supply 800 validation frames, and 65–70 supply 600 synthetic test frames. Videos 71–80 are reserved for a separate real-image diagnostic set. B1–B4 models are evaluated on the same four synthesis conditions. We report the official 12k comparison and its 50k continuation separately. External comparisons cover Restormer, NAFNet, and HINet (Zamir et al., 2022; Chen et al.,



Figure 2: **Evaluation datasets for surgical smoke restoration.** SurgiMist-Cholec and LSD3K provide exact synthetic image pairs; DesmokeData provides real pairs with motion-compensated references; DeSmoke-LAP and CholecOOD contain real-smoke images without per-frame clean targets. The figure summarizes the fixed restoration benchmark; the controlled perception study additionally uses CholecSeg8k labels and three surgical RGB sources.

2022; 2021), with LSD3K (Chang & Chen, 2024) as the alternative training distribution. Paired tests use LSD3K (200 images) and DesmokeData (Xia et al., 2024) (961 pairs). Real-image diagnostics use DeSmoke-LAP (Pan et al., 2022) (500 frames) and the held-out Cholec80 set (CholecOOD, 215 frames). We retain full matrices, a 2,520-example count-matched comparison, and four scaling protocols in Appendix C.

Controlled surgical study. RGB sources are Cholec80, BernBypass70, and StrasBypass70 (Lavanchy et al., 2024). We select 9,020 frames from 22,000 decoded frames. Each single-source restoration run uses 2,400 distinct training images; a pooled run uses 800 images per source. Restoration training uses H, MCBM, and full B4, including its small coordinate displacement. All three single-source training distributions are evaluated on all three target sources. CholecSeg8k (Hong et al., 2020) supplies semantic labels derived from Cholec80, and is not counted as an independent fourth surgical source. Downstream validation/test videos are excluded from Cholec80 restoration training.

The formal semantic test panel contains 96 frames, 32 from each of three held-out videos (18, 25, and 43). We compute 13-class mIoU from the pooled pixel confusion matrix, assigning zero to a class with zero union. Training seeds are 3407, 3408, and 3409. Reported uncertainties are sample standard deviations across these three training seeds, not confidence intervals over patients or hospitals.

Training and comparisons. The controlled study trains NAFNet restorers from scratch with an ℓ_1 objective and DeepLabV3–ResNet50 predictors (Chen et al., 2017) from COCO-pretrained weights with a new 13-class head. Each run attempts 6,000 batches of four 256×256 crops with AdamW, initial learning rate 2×10^{-4}, cosine decay, and gradient clipping at norm 1; augmented runs include 10% original-image exposure. Mixed-precision skipped updates are recorded; the budget describes attempted batches. Final checkpoints are used for reporting. Table 3 compares seven segmentation training strategies on a balanced panel of five test renderers (H, MCBM, H+Gaussian, B4-photo without Gaussian, and B4-photo) at three severities (15 synthetic conditions), with original images reported separately. We first compute each condition's pooled-confusion mIoU and then average the 15 scores equally within seed. These component studies followed inspection of the initial four-strategy results and are exploratory.

Metrics and processing routes. Restoration is evaluated by PSNR, SSIM (Wang et al., 2004), and, in the fixed benchmark, LPIPS (Zhang et al., 2018). Real-image diagnostics include NIQE, BRISQUE, and FADE (Mittal et al., 2013; 2012; Choi et al., 2015), plus descriptive contrast, sharpness, edge, entropy, and saturation statistics. They are proxies for appearance, not clinical fidelity. We compare a predictor trained directly with smoke augmentation against an original-trained predictor preceded by a smoke-trained restorer. Independent forward-pass timing on A5000 and A6000 GPUs characterizes architecture cost at batch one and 288×512 resolution; it excludes decoding, transfers, and display.

Blinded doctor review. Five doctors each compared B4- and LSD3K-trained Restormer outputs for the same 100 real-smoke images: 50 from DeSmoke-LAP and 50 from Cholec80 RealSmoke

Table 1: Restormer recovery averaged equally over the four B1–B4 test conditions at the official 12k budget.

Training	PSNR ↑	SSIM ↑	LPIPS ↓
B1	26.52	0.9101	0.0971
B2	25.96	0.8846	0.1027
B3	26.78	0.9037	0.0945
B4	**27.94**	**0.9303**	**0.0867**

Each configuration is evaluated on the same four test conditions. Bold indicates the best mean. Full cross-condition results and checkpoint details are reported in the appendix.

Table 2: Real-smoke restoration diagnostics on DeSmoke-LAP. Lower NIQE, BRISQUE, and FADE are better. B4 has the best value for each metric across the shown training variants within all three backbones.

Backbone	Training	NIQE ↓	BRISQUE ↓	FADE ↓
Unrestored input		6.7719	41.81	1.1111
Restormer	B0	5.6024	30.53	0.4831
	B4	**5.1997**	**24.79**	**0.4550**
	LSD3K	5.3048	28.94	0.5272
NAFNet	B0	5.6694	34.03	0.5189
	B4	**4.8011**	**24.68**	**0.4419**
	LSD3K	5.6206	29.69	0.5654
HINet	B1	5.8255	31.62	0.5289
	B4	**5.1546**	**27.59**	**0.4660**
	LSD3K	5.5728	31.65	0.5959

B0 is the homogeneous baseline and B1 the heterogeneous one-layer control. B4 and LSD3K are compared across all three backbones; full diagnostics and Cholec80-OOD results are in the appendix.

OOD. Output order was balanced, and each image pair was rated by all five doctors for visibility, structure preservation, introduced artifacts, and offline viewing preference.

5 Results

5.1 Restoration quality and external behavior

The full generator improves the fixed benchmark average. In the 12k benchmark comparison, B4 achieves 27.94 dB PSNR, 0.9303 SSIM, and 0.0867 LPIPS averaged over the same four test conditions (Table 1). These are the best means among B1–B4. The condition-level matrix shows that the result is broader than the matched B4 condition: B4 has the best PSNR and SSIM on all four test conditions, while the lowest LPIPS varies by condition. Continuing the benchmark B4 checkpoint to the cumulative 50k budget raises its own mean PSNR to 30.82 dB and improves SSIM and LPIPS to 0.9568 and 0.0511. At that budget, B3 leads mean PSNR (31.05 dB), and B4 leads the other two mean metrics. Thus, adding the complete image effects provides a useful training distribution, with the ranking depending on metric and optimization budget.

Real-image diagnostics favor B4 on DeSmoke-LAP across backbones. B4-trained Restormer, NAFNet, and HINet each have lower NIQE, BRISQUE, and FADE than their LSD3K-trained counterpart on DeSmoke-LAP (Table 2). This consistency across architectures supports the utility of the synthetic training distribution for that real-image set. External behavior is not uniform: LSD3K-trained models win all three paired fidelity metrics on both paired test datasets in this comparison, and CholecOOD gives mixed no-reference rankings. We therefore interpret the DeSmoke-LAP result as a real-image appearance diagnostic, rather than proof that B4 recovers the most faithful

Table 3: Segmentation mIoU (%) on original images and the balanced 15-condition synthetic panel. Entries show mean ± reported SD. Δ is the difference between the synthetic means and the Original-only mean. H+Gaussian uses the repeated balanced-panel result. Blur-only is an exploratory control.

Training augmentation	Original images	Synthetic panel	Δ (pp)
Original-only	**50.26 ± 0.70**	34.47 ± 0.72	–
Blur-only (expl.)	49.48 ± 0.88	34.62 ± 3.07	+0.15
H	47.78 ± 1.66	48.76 ± 2.23	+14.29
MCBM	49.26 ± 1.49	49.78 ± 1.55	+15.31
H + blur	49.38 ± 1.39	49.21 ± 1.32	+14.74
B4-photo, no blur	49.51 ± 0.53	49.64 ± 0.27	+15.17
B4-photo	49.50 ± 0.17	**50.18 ± 0.12**	**+15.71**

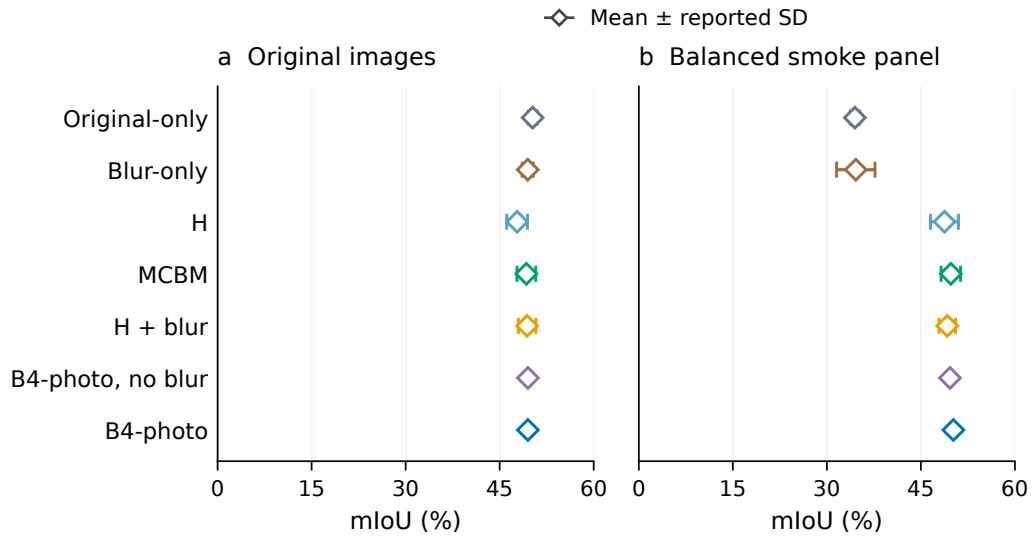


Figure 3: **Smoke augmentation improves the balanced synthetic panel while preserving performance on original images.** Each point shows mean mIoU and reported SD for seven training strategies on the original-image panel and the equal-weight 15-condition synthetic panel. The H+Gaussian point uses the repeated balanced-panel estimate reported in Table 3.

anatomy. Complete paired values, all eight real-image statistics, count-matched tests, and data-scale results are retained in the appendix.

5.2 Smoke-robust semantic perception

Controlled augmentation closes a large synthetic robustness gap. B4-photo reaches $50.18 \pm 0.12\%$ mIoU on the balanced 15-condition panel, compared with $34.47 \pm 0.72\%$ for original-only training (Table 3). The paired mean gain is 15.71 percentage points. At severe B4-photo smoke ($\bar{T} = 0.35$), the corresponding scores are 50.28% and 19.42%. Figure 3 compares original-image performance with the balanced synthetic panel; severity-resolved profiles are provided in Appendix B. Original-image mIoU remains 49.50%, compared with 50.26% for original-only training, exposing the trade-off between original appearance and synthetic robustness.

Component controls identify a shared augmentation benefit. B4-photo exceeds H by 1.42 percentage points and matched MCBM by 0.41 points in the balanced mean. The repeated H+Gaussian balanced score is $49.21 \pm 1.32\%$, compared with $50.18 \pm 0.12\%$ for B4-photo and $49.64 \pm 0.27\%$

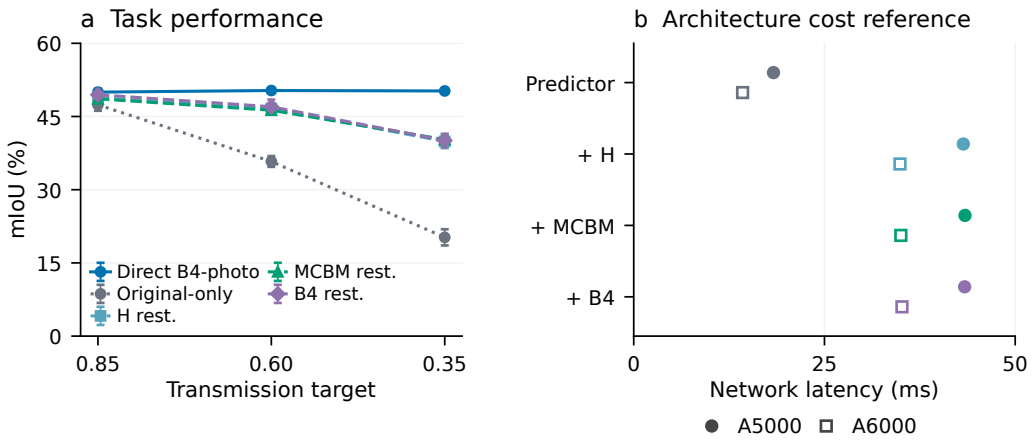


Figure 4: **Direct augmentation improves severe-smoke task accuracy; a restoration front end adds network cost.** Left: direct B4-photo augmentation and restoration followed by an original-trained predictor, averaged equally over all five test renderers at each severity. Right: independently measured network-forward latency at batch one. Cost bars characterize the two architectures and do not constitute an end-to-end clinical throughput measurement.

for B4-photo without Gaussian smoothing. Gaussian-only training scores 34.62%, close to the original-only baseline. B4-photo has the highest synthetic mean among these component controls, while the Gaussian-only comparison shows that ordinary blur alone does not reproduce the smoke-augmentation gain. B4-photo's small observed seed spread is encouraging, but three seeds and three test videos do not establish population-level reliability.

5.3 Choosing a task processing route

Direct augmentation outperforms the tested restoration front ends under heavy smoke. The balanced synthetic mean is 50.18% for direct B4-photo training and 45.52% for B4 restoration followed by prediction. On B4-photo test images at $\bar{T} = 0.35$, direct training yields 50.28% mIoU. The three tested restoration front ends yield 39.24% (H), 39.54% (MCBM), and 39.45% (B4) before the same original-trained predictor. These results hold the downstream test frames and original predictor fixed for the restoration routes. They show that the restoration networks recover part of the lost task performance, while direct task adaptation is more effective under the tested degradation. On original images, restoration routes score approximately 50.6%, compared with 49.5% for direct B4-photo training; the route advantage is condition-dependent.

Image fidelity and task accuracy need separate validation. The PSNR-leading restorer and the mIoU-leading restorer differ in 10 of 15 matched source–condition groups (three restoration training sources and five initial synthetic test conditions, after averaging seeds). The average test-oracle task gap is small (0.26 percentage points, maximum 0.70), so the mismatch should not be interpreted as wholesale failure of fidelity metrics. In the three-source study, B4 minus H restoration PSNR is also close to zero on average at common transmission 0.60 (-0.004 dB within source and $+0.008$ dB across sources). The task-level gain of augmentation is therefore a separate result from a restoration-generalization claim.

The simpler task route also avoids a restoration network. At 288×512, the predictor architecture takes 18.32 ms on A5000 and 14.27 ms on A6000. A B4-trained restoration network followed by the predictor takes 43.38 ms and 35.16 ms, respectively. Peak allocated memory rises from 365 to 432 MiB. These measurements use identical input dimensions and precision settings; the predictor-only cost reference uses the original-trained head. They support a network-level efficiency

Table 4: Blinded paired judgments of B4- and LSD3K-trained Restormer outputs on 100 real-smoke images. Values are counts (% of 500 judgments). Five doctors each rated all 100 image pairs; model order was balanced.

Criterion	B4 selected	LSD3K selected	Tie
Visibility	218 (43.6%)	214 (42.8%)	68 (13.6%)
Structure preservation	210 (42.0%)	206 (41.2%)	84 (16.8%)
Offline viewing preference	217 (43.4%)	212 (42.4%)	71 (14.2%)
Introduced artifact flagged	35 (7.0%)	42 (8.4%)	–

Each image was rated by all five doctors. Artifact counts are per model output. These are paired judgments on repeated images, not 500 independent images.

advantage for direct prediction, while quality comparisons come from the separately evaluated augmented heads.

5.4 Blinded doctor review

Across 500 paired judgments, doctors selected B4 slightly more often than LSD3K for visibility, structure preservation, and offline viewing preference (Table 4). The differences are small and ties account for 14–17% of responses. Doctors flagged introduced artifacts in 35 B4 outputs (7.0%) and 42 LSD3K outputs (8.4%). The comparison assesses restoration preferences on real-smoke inputs; it does not test whether doctors can distinguish synthetic images from naturally smoky images.

6 Conclusion and limitations

SurgiMist connects controllable smoke synthesis to restoration benchmarking and task-oriented training. The complete generator performs strongly on the fixed restoration benchmark and lowers NIQE, BRISQUE, and FADE on DeSmoke-LAP across three backbones. In the controlled perception study, smoke augmentation produces a large robustness gain, and direct task training outperforms the tested restoration front ends under severe synthetic smoke. Simpler optical controls recover much of the same segmentation benefit, making the framework's control and evaluation interface central to its value.

The current evidence is limited by synthetic downstream corruptions, three held-out semantic videos, relative rather than metric depth, and image-space approximations to scattering. Source images are not verified to be naturally smoke-free. Natural-smoke downstream annotations are unavailable. The blinded review compares restoration outputs on real-smoke images and does not measure synthetic-versus-natural source recognition. The next validation step is a prospectively specified evaluation on naturally smoky, independently annotated surgical sequences, including structure preservation and clinical task relevance.

Reproducibility statement

The appendix specifies the source splits, rendering equations, training budgets, metric aggregation, and timing protocol. The linked anonymous repository provides the smoke-synthesis code and benchmark release. The manuscript source package contains the LaTeX source, figures, and tables. External datasets remain subject to their original terms.

Ethics statement

This work studies image synthesis and algorithmic evaluation on existing surgical video datasets. Synthetic images and no-reference image-quality scores do not establish diagnostic safety or preservation of clinically meaningful structures. The evaluated systems are research prototypes and have not been validated for clinical deployment. No new patient recruitment occurred; the blinded review uses existing image outputs.

REFERENCES

Borja Bovcon, Jon Muhovič, Janez Pers, and Matej Kristan. The MaSTr1325 dataset for training deep USV obstacle detection models. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019. URL https://vicos.si/publications/bovcon2019the/.

Gabriel J. Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009. doi: 10.1016/j.patrec.2008.04.005. URL https://www.sciencedirect.com/science/article/pii/S0167865508001220.

Wenhui Chang and Hongming Chen. LSD3K: A benchmark for smoke removal from laparoscopic surgery images. *arXiv preprint arXiv:2407.13132*, 2024. URL https://arxiv.org/abs/2407.13132.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. URL https://arxiv.org/abs/1706.05587.

Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. HINet: Half instance normalization network for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 182–192, 2021. doi: 10.1109/CVPRW53098.2021.00027. URL https://openaccess.thecvf.com/content/CVPR2021W/NTIRE/papers/Chen_HINet_Half_Instance_Normalization_Network_for_Image_Restoration_CVPRW_2021_paper.pdf.

Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *Computer Vision – ECCV 2022*, volume 13667 of *Lecture Notes in Computer Science*, pp. 17–33. Springer, 2022. doi: 10.1007/978-3-031-20071-7_2. URL https://link.springer.com/chapter/10.1007/978-3-031-20071-7_2.

Lark Kwon Choi, Jaehee You, and Alan Conrad Bovik. Referenceless prediction of perceptual fog density and perceptual image defogging. *IEEE Transactions on Image Processing*, 24(11):3888–3901, 2015. doi: 10.1109/TIP.2015.2456502. URL https://live.ece.utexas.edu/publications/2015/LarkKwonChoi_TIP_2015_Referenceless_Prediction_of_Perceptual_Fog_Density_and_Perceptual_Image_Defogging.pdf.

Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2341–2353, 2011. doi: 10.1109/TPAMI.2010.168. URL https://doi.org/10.1109/TPAMI.2010.168.

W.-Y. Hong, C.-L. Kao, Y.-H. Kuo, J.-R. Wang, W.-L. Chang, and C.-S. Shih. CholecSeg8k: A semantic segmentation dataset for laparoscopic cholecystectomy based on Cholec80. *arXiv preprint arXiv:2012.12453*, 2020. URL https://arxiv.org/abs/2012.12453.

Joël L. Lavanchy, Sanat Ramesh, Diego Dall'Alba, Cristians Gonzalez, Paolo Fiorini, Beat P. Müller-Stich, Philipp C. Nett, Jacques Marescaux, Didier Mutter, and Nicolas Padoy. Challenges in multi-centric generalization: Phase and step recognition in Roux-en-Y gastric bypass surgery. *International Journal of Computer Assisted Radiology and Surgery*, 19(11):2249–2257, 2024. doi: 10.1007/s11548-024-03166-3. URL https://link.springer.com/article/10.1007/s11548-024-03166-3.

Anish Mittal, Anush K. Moorthy, and Alan C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. doi: 10.1109/TIP.2012.2214050. URL https://doi.org/10.1109/TIP.2012.2214050.

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. doi: 10.1109/LSP.2012.2227726. URL https://doi.org/10.1109/LSP.2012.2227726.

Yirou Pan, Sophia Bano, Francisco Vasconcelos, Hyun Park, Taikyeong Ted. Jeong, and Danail Stoyanov. DeSmoke-LAP: Improved unpaired image-to-image translation for desmoking in laparoscopic surgery. *International Journal of Computer Assisted Radiology and Surgery*, 17(5): 885–893, 2022. doi: 10.1007/s11548-022-02595-2. URL https://link.springer.com/article/10.1007/s11548-022-02595-2.

Jialun Pei, Diandian Guo, Donghui Yang, Zhixi Li, Yuxin Feng, Long Ma, Bo Du, and Pheng-Ann Heng. Benchmarking endoscopic surgical image restoration and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37413–37422, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Pei_Benchmarking_Endoscopic_Surgical_Image_Restoration_and_Beyond_CVPR_2026_paper.html.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, 2018. doi: 10.1007/s11263-018-1072-8. URL https://link.springer.com/article/10.1007/s11263-018-1072-8.

Christos Sakaridis, Dengxin Dai, and Luc Van Gool. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10765–10775, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Sakaridis_ACDC_The_Adverse_Conditions_Dataset_With_Correspondences_for_Semantic_Driving_ICCV_2021_paper.html.

Junseong Shin, Seungwoo Chung, Yunjeong Yang, and Tae Hyun Kim. HazeFlow: Revisit haze physical model as ODE and non-homogeneous haze generation for real-world dehazing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6263–6272, 2025. URL https://www.openaccess.thecvf.com/content/ICCV2025/papers/Shin_HazeFlow_Revisit_Haze_Physical_Model_as_ODE_and_Non-Homogeneous_Haze_ICCV_2025_paper.pdf.

Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *IEEE Transactions on Image Processing*, 32:1927–1941, 2023. doi: 10.1109/TIP.2023.3256763. URL https://doi.org/10.1109/TIP.2023.3256763.

Andru Putra Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy. EndoNet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging*, 36(1):86–97, 2017. doi: 10.1109/TMI.2016.2593957. URL https://doi.org/10.1109/TMI.2016.2593957.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861. URL https://www.colorado.edu/lab/live/publications/zwang_ssim_ieeeip2004.pdf.

Wenyao Xia, Victoria Fan, Terry Peters, and Elvis C. S. Chen. A new benchmark in vivo paired dataset for laparoscopic image de-smoking. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15001 of *Lecture Notes in Computer Science*, pp. 3–13. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72378-0_1. URL https://papers.miccai.org/miccai-2024/022-Paper1213.html.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2. In *Advances in Neural Information Processing Systems*, volume 37, pp. 21875–21911. Curran Associates, Inc., 2024. doi: 10.52202/079017-0688. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/26cfdcd8fe6fd75cc53e92963a656c58-Abstract-Conference.html.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5728–5739, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/

html/Zamir_Restormer_Efficient_Transformer_for_High-Resolution_Image_Restoration_CVPR_2022_paper.html.

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*, pp. 586–595, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.