

Trustworthy Endoscopic Super-Resolution

Julio Silva-Rodríguez and Ender Konukoglu

Computer Vision Lab, ETH Zurich, Zurich, Switzerland
jusilva@ethz.ch

Abstract. Super-resolution (SR) models are attracting growing interest for enhancing minimally invasive surgery and diagnostic videos under hardware constraints. However, valid concerns remain regarding the introduction of hallucinated structures and amplified noise, limiting their reliability in safety-critical settings. We propose a direct and practical framework to make SR systems more trustworthy by identifying where reconstructions are likely to fail. Our approach integrates a lightweight error-prediction network that operates on intermediate representations to estimate pixel-wise reconstruction error. The module is computationally efficient and low-latency, making it suitable for real-time deployment. We convert these predictions into operational failure decisions by constructing Conformal Failure Masks (CFM), which localize regions where the SR output should not be trusted. Built on conformal risk control principles, our method provides theoretical guarantees for controlling both the tolerated error limit and the miscoverage in detected failures. We evaluate our approach on image and video SR, demonstrating its effectiveness in detecting unreliable reconstructions in endoscopic and robotic surgery settings. To our knowledge, this is the first study to provide a model-agnostic, theoretically grounded approach to improving the safety of real-time endoscopic image SR. Code is available: [Endoscopic-CFM](#).

Keywords: Endoscopy · Super-resolution · Trustworthy AI.

1 Introduction

Minimally invasive surgery and diagnostic procedures, including laparoscopic and robot-assisted techniques, have advanced surgical practice by reducing recovery time, tissue trauma, and scarring [32]. In this context, high-quality imaging is essential for surgical safety, precise dissection, and identification of critical anatomical structures [19]. However, acquiring high-resolution images in clinical practice remains challenging due to hardware constraints, artificial noise sources, and patient comfort requirements, particularly in real-time clinical settings [14].

Recent advances in machine learning aim to overcome these limitations with super-resolution (SR) deep neural networks that recover fine details and reduce noise. Lightweight models for real-time single-image [27,36] and video [10,11] SR have shown promising results, including for endoscopic data [28]. Nevertheless, SR models, and generative models in general, raise concerns about reliability, potentially producing unnoticed hallucinations and amplifying artifacts. In this

context, medical SR approaches must provide not only accurate reconstructions but also robust estimates of failure awareness to ensure safe clinical use.

Pitfalls of current uncertainty quantification frameworks. Failure detection has often been approached through the lens of uncertainty estimation, under the assumption that uncertain predictions correlate with high error. In image classification, softmax scores often correlate with correct class predictions [22] and can be further calibrated [21]. However, these approaches do not directly extend to continuous regression outputs, as in image SR. Other alternatives leverage stochastic processes (e.g., Monte Carlo dropout [18] or diffusion-based models [1]) or ensembles [25] to estimate predictive variability as a proxy for uncertainty. While effective, such methods typically require multiple forward passes or architectural modifications, making them less suitable for real-time pipelines. A lighter alternative is to learn failure estimates directly with auxiliary networks. These approaches have been explored mainly in discriminative contexts, such as classification and segmentation [15,16] or multi-modal understanding [24], where reliability can be tied to well-defined probabilistic scores, e.g., true-class probabilities. Extending these ideas to image SR is less straightforward. In SR, outputs are continuous and structured, and failure must be defined relative to task-specific thresholds rather than an intrinsic notion of correctness. Crucially, beyond these limitations, none of these alternatives alone yield operational failure decisions with robust, distribution-free guarantees suitable for clinical use. Based on all the above, we find that existing uncertainty methods are not directly applicable under the constraints of using existing real-time SR models.

Conformal prediction (CP) [2,47] is a framework that aims to provide model-agnostic, distribution-free, and theoretically valid guarantees on machine learning models’ outputs. In its most widely adopted version, split conformal prediction [37,47], a calibration subset is used to post-process neural network outputs into operational decisions, e.g., decision intervals for regression [38,43,4] or class sets for classification [33,3], that guarantee a given coverage of correctness. Despite growing interest, CP remains relatively underexplored in medical imaging generation and reconstruction, although recent work has shown promising results in image classification [30,12], grading [31], segmentation [9,7,35], multi-modal models [41,42], and surgical instrument forecasting [39], among other tasks [40,45,13,26]. In the computer vision community, CP for image-to-image generation has been typically addressed by producing interval-valued outputs [4,44] instead of single images. In such a pipeline, the reliability is communicated qualitatively to the practitioner via the interval size; e.g., pixels with larger intervals are likely to be less reliable. However, as recently shown in [1], these heatmaps are hardly interpretable and do not yield fully operational failure decisions. In contrast, the very recent work in [1], which is more closely aligned with our efforts, explores the creation of binary masks that explicitly highlight image reconstruction errors.

Contributions. In this work, we address the challenge of quantifying the reliability of SR models, focusing on an application in endoscopic imaging. Specifically, we propose a novel bi-level conformal risk control procedure that allows

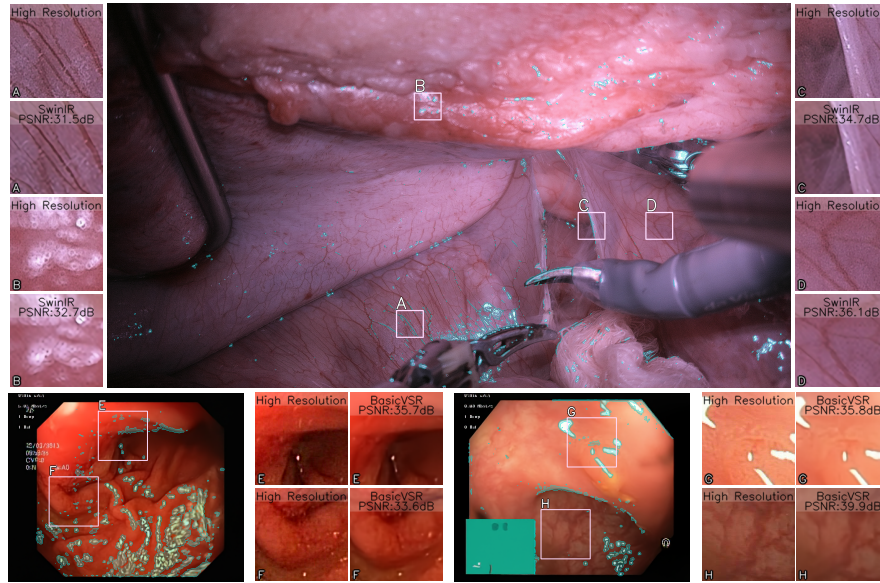


Fig. 1: **Conformal Failure Masks for trustworthy SR.** These (*blue-ish*) provide operational detection of hallucinations over the high-resolution output. Masks are created at an error level of 22 dB and tolerating a false-negative rate of 5%. The operational values can be adjusted, maintaining theoretical guarantees.

the user to control both the permissible level of reconstruction error and miscoverage on detected failures. Technically, we first propose learning a pixel-level error estimate using intermediate representations from a frozen SR model, with a lightweight module, termed the *Reconstruction Error Network*. We then propose creating operational *Conformal Failure Masks* (CFMs, see Fig. 1, *blue-ish masks*) to identify generated regions that practitioners should not trust. For this purpose, we introduce the concept of *reconstruction failure levels* to create binary failure-detection sub-tasks (e.g., as measured by PSNR). Our CFMs are then created by rejecting pixels whose estimated errors exceed a threshold. Specifically, the operative rejection threshold is chosen to control the false-negative rate at the target failure level by building on well-established theoretical guarantees for conformal risk control [5]. The proposed procedure does not aim to correct failed pixels, but to serve as a safety overlay. Masked regions indicate where SR should not be trusted: users may inspect the raw image, reduce reliance on SR, or spend more time assessing suspicious/noisy structures. We present extensive experiments demonstrating that our framework can successfully measure and control the reconstruction error across three tasks and SR models. By controlling both the desired failure level and miscoverage, practitioners can flexibly generate practical reliability masks to support more informed clinical decisions and greater trust in technology.

2 Methods

Image SR. Let $\mathcal{X} \subset \mathbb{R}^{\Omega \times C}$ denote the low-resolution (LR) image space, where $\Omega \subset \mathbb{Z}^2$ is the discrete spatial domain, $\Omega = \{0, \dots, W-1\} \times \{0, \dots, H-1\}$, and C denotes the number of channels. We define the high-resolution (HR) image space $\mathcal{Y} \subset \mathbb{R}^{\Omega' \times C}$, where, given an upscaling factor $s \in \mathbb{N}, s > 1$, the HR spatial domain is $\Omega' = \{0, \dots, sW-1\} \times \{0, \dots, sH-1\}$, so that $|\Omega'| = s^2|\Omega|$. Super-resolution (SR) aims to recover an HR image from its LR observation by learning a mapping $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$, parameterized by a deep neural network with parameters $\theta = \{\theta_{\text{enc}}, \theta_{\text{rec}}\}$. The network is typically composed of two stages: a feature extraction (encoder/bottleneck) module $f_{\theta_{\text{enc}}} : \mathcal{X} \rightarrow \mathcal{F}$, which maps the input image into a deep latent feature space $\mathcal{F} \subset \mathbb{R}^{\Omega \times F}$ with F feature channels, and a reconstruction (upsampling) module $f_{\theta_{\text{rec}}} : \mathcal{F} \rightarrow \mathcal{Y}$, which performs spatial upsampling followed by nonlinear transformations to reconstruct the HR image.

2.1 Lightweight SR error quantification

We first aim to obtain a measure of error for pixel-level predictions given a pre-trained SR model. For that purpose, we propose a data-driven learning procedure in which an auxiliary module projects the intermediate feature representations of the SR encoder into a predictive unreliability measure, i.e., $f_\Phi : \mathcal{F} \rightarrow \hat{\mathcal{E}}$, where $\hat{\mathcal{E}} \subset \mathbb{R}^{\Omega' \times 1}$. The function f_Φ , parameterized by a deep neural network Φ and coined the *Reconstruction Error Network*, is fitted to produce a continuous response, with larger values expected to align with higher reconstruction error.

Optimization. Let $\mathcal{D}_{\text{train}} = \{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i \in \mathcal{T}}$, with $\mathcal{T} = \{i \in \mathbb{N} : 0 \leq i \leq I-1\}$, such that $\mathbf{X}_i \in \mathcal{X}$ and $\mathbf{Y}_i \in \mathcal{Y}$ are paired LR and HR images, denote a small training dataset with fresh data points, not used to pre-train the SR model. Also, let $\hat{\mathbf{Y}}_i = f_\theta(\mathbf{X}_i)$ denote the reconstructed image and $\mathbf{F}_i = f_{\theta_{\text{enc}}}(\mathbf{X}_i)$ its intermediate feature representation. The network is optimized via mini-batch gradient descent to approximate the pixel-level squared error between predicted and reference HR images from the intermediate features:

$$\min_{\Phi} \frac{1}{|\mathcal{T}||\Omega'|} \sum_{i \in \mathcal{T}} \sum_{\Omega'} \left(\frac{1}{C} \sum_{c=0}^{C-1} (\hat{\mathbf{Y}}_i - \mathbf{Y}_i)^2 - f_\Phi(\mathbf{F}_i) \right)^2. \quad (1)$$

Architecture. Bicubic interpolation is used for initial upsampling, followed by a sequence of convolutional blocks and a final linear layer. Each block contains standard operations, i.e., convolution, ReLU, and batch normalization. This simple design provides a lightweight solution that can run in parallel with the SR model’s reconstruction block without introducing additional latency.

2.2 Conformal Failure Masks

We now propose transforming the pixel-level continuous error scores into operational binary masks, referred to as *Conformal Failure Masks* (CFMs).

Reconstruction failure levels. Unlike standard classification tasks, errors in image reconstruction are continuous, reflecting varying degrees of incorrectness depending on the difficulty of the task. To enable binary decisions (failure / non-failure) from continuous errors, we first define a per-pixel observed error map $\mathbf{E}_i \in \mathbb{R}^{\Omega' \times 1}$ between the predicted and reference HR images, e.g., using peak signal-to-noise ratio (PSNR) or squared difference. In practice, we adopt PSNR, as is standard in SR [10, 11, 28, 1]. Then, for a given failure level $\tau^{\text{fail}} \geq 0$ (e.g., PSNR = 22 dB), the target binary mask is defined as $\mathbf{M}_i^{\tau^{\text{fail}}} = \mathbf{1}\{\mathbf{E}_i^{\text{PSNR}} \leq \tau^{\text{fail}}\} \in \{0, 1\}^{\Omega' \times 1}$, where $\mathbf{1}\{\cdot\}$ is the indicator function. This procedure converts reconstruction errors into a set of binary classification sub-tasks, each corresponding to a user-controlled specific failure level, τ^{fail} .

Bi-level risk control. Next, given the estimate of our error network for a particular image, $\hat{\mathbf{E}}_i \in \hat{\mathcal{E}}$, these scores are binarized using an operative threshold, τ , to approximate its oracle failure mask at a given failure level, $\mathbf{M}_i^{\tau^{\text{fail}}}$, such that $\hat{\mathbf{M}}_i^\tau = \mathbf{1}\{\hat{\mathbf{E}}_i \geq \tau\} \in \{0, 1\}^{\Omega' \times 1}$. However, due to imperfect error estimation, the operative threshold must balance a suitable risk measure. In clinical settings, a user may wish to control the number of false negatives (SR hallucinations or noisy regions that are not flagged), that is, the false-negative-rate (FNR), so that it does not exceed a predefined frequency, $\alpha \in [0, 1]$. To do so, we propose a conformal procedure to establish the operative threshold that theoretically guarantees a miscoverage rate (α) for an arbitrary failure level (τ^{fail}):

$$\text{FNR}(\tau^{\text{fail}}, \tilde{\tau}) \leq \alpha. \quad (2)$$

Let us assume access to a set of LR and HR images different from those used to train the SR and error networks (the calibration set in split conformal prediction), $\mathcal{D}_{\text{cal}} = \{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i \in \mathcal{C}}$, with $\mathcal{C} = \{i \in \mathbb{N} : I \leq i \leq I + M - 1\}$. The false-negative rate across the entire pixel population is computed as:

$$\text{FNR}_{\mathcal{D}_{\text{cal}}}(\tau^{\text{fail}}, \tau) = \frac{1}{N} \sum_{i \in \mathcal{C}} \sum_{\Omega'} \mathbf{1}\{\mathbf{M}_i^{\tau^{\text{fail}}} = 1 \wedge \hat{\mathbf{M}}_i^\tau = 0\}, \quad (3)$$

where $N = \sum_{i \in \mathcal{C}} \sum_{\Omega'} \mathbf{1}\{\mathbf{M}_i^{\tau^{\text{fail}}} = 1\}$ corresponds to the actual positive pixels, i.e., pixels with ground truth error above the target failure level. For a desired miscoverage, α , and failure level, τ^{fail} , the operative threshold is:

$$\tilde{\tau} = \sup \left\{ \tau : \frac{N}{N+1} \text{FNR}_{\mathcal{D}_{\text{cal}}}(\tau^{\text{fail}}, \tau) - \frac{1}{N+1} \leq \alpha \right\}. \quad (4)$$

Since $\text{FNR}(\tau^{\text{fail}}, \tau)$ is monotone in τ , and analogously to standard conformal risk control proofs [5], the bi-level finite-sample guarantee in Eq. 2 will hold for new data that are, at least, exchangeable with the calibration distribution.

3 Experiments

3.1 Setup

Datasets. The proposed methods are validated on two different types of endoscopy data: robotic surgery imagery and minimally-invasive diagnostic videos.

For the first scenario, we employ SurgiSR4K [23], a super-resolution-focused dataset consisting of images at three resolution levels, up to native 4K. Two scenarios are contemplated: a $2\times$ image enhancement, from 480x270p to 960x540p (**Surgi-2 $\times$**), and a more challenging $4\times$ super-resolution, from 960x540p up to 3840x2160p (**Surgi-4 $\times$**). The samples are randomly split into training (10 videos, 335 images) and test (15 videos, 465 images) subsets. For the second scenario, HyperKvasir [8] (**HKvasir**), which comprises endoscopic videos from gastrointestinal tract diagnostic procedures, is used. Specifically, we follow prior work [28] and investigate a $4\times$ SR enhancement with noise injection, from 180x144p to 720x576p images. Also, the same data splits as in [28] are used. Training data in [28] are used to pre-train a video SR model (236 videos, 11,800 images), while the validation (14 videos, 700 images) and test (27 videos, 1,350 images) subsets are used to train and evaluate our failure detection framework.

SR models. First, for the Surgi-s $\times$ datasets, we use SwinIR [27], a transformer-based single-image SR model trained on natural images. Specifically, we use the default weights of its lightweight versions (**SwinIR-L-s $\times$**), which are designed for real-time SR applications (0.9M parameters; $\simeq 30$ ms/image¹). Second, for HKvasir, we focus on video SR models, which propagate temporal features to improve consistency. Specifically, we used the lightweight **BasicVSR** [10] (2.9M parameters; $\simeq 20$ ms/image). In our real-time scenario (only forward feature propagation allowed), we observed performance comparable to that of more recent methods, e.g., BasicVSR⁺⁺ [11] or MedVSR [28]; therefore, we prioritized simplicity. Training hyper-parameters were as in the recent benchmark in [28].

Reconstruction Error Network. The network comprises 2 blocks, each with 64 features (70K parameters; latency negligible relative to SR). Using the small training split, as described in Section 2.1, a 30-epoch training run with a batch size of 1 video, the Adam optimizer, and a cosine-decay scheduler with a base learning rate of 10^{-3} , is performed. All details remain fixed across tasks.

Conformal Failure Masks (CFM) are created as described in Section 2.2, at three target failure levels using cutoff (τ^{fail}) PSNR values of 22, 24, and 26 dB; and two FNR miscoverage, $\alpha \in \{0.10, 0.05\}$. Unless otherwise specified, exploratory experiments are conducted with a baseline failure level of 22 dB.

Metrics and evaluation protocol. Standard metrics for misclassification detection are used to assess the performance of our Reconstruction Error Network [15,20,24], i.e., area under the ROC curve (AUROC) and False Positive Rate at 95% True Positive Rate (FPR95). CFMs are assessed by measuring FNR at the target failure level. Also, the average mask size per image and PSNR in the non-rejected regions, as in [1], are obtained. The procedure is repeated and evaluated on 100 random splits of the test videos (70% – 30% for $\mathcal{D}_{\text{cal}}$ and test).

3.2 Results

Main results. Fig. 2 confirms the positive outcome of our Reconstruction Error Network to discriminate erroneous reconstructions in the three use-cases, pro-

¹ Latency metrics are computed on a NVIDIA RTX A4000 16GB GPU workstation.

Table 1: Quantitative evaluation of our Conformal Failure Masks.

SR Method		$\alpha = 0.10$			$\alpha = 0.05$		
		Avg. PSNR	FNR (%)	Avg. Mask Size (%)	Avg. PSNR	FNR (%)	Avg. Mask size (%)
Surgi-2 $\times$	Bicubic	36.5 ± 1.2	–	–	36.5 ± 1.2	–	–
	SwinIR-L-2 $\times$	38.7 ± 1.5	–	–	38.7 ± 1.5	–	–
	+CFM ($\tau^{\text{fail}} = 22$ dB)	40.9 ± 0.8	10.8	2.7 ± 1.3	41.4 ± 0.7	5.3	4.0 ± 1.7
	+CFM ($\tau^{\text{fail}} = 24$ dB)	41.3 ± 0.7	10.5	3.6 ± 1.6	41.8 ± 0.7	5.1	5.4 ± 2.1
	+CFM ($\tau^{\text{fail}} = 26$ dB)	41.7 ± 0.7	10.4	4.8 ± 2.0	42.3 ± 0.7	5.0	7.4 ± 3.1
Surgi-4 $\times$	Bicubic	33.0 ± 1.8	–	–	33.0 ± 1.8	–	–
	SwinIR-L-4 $\times$	33.7 ± 1.9	–	–	33.7 ± 1.9	–	–
	+CFM ($\tau^{\text{fail}} = 22$ dB)	34.6 ± 1.2	9.6	4.0 ± 4.2	34.7 ± 1.2	4.6	5.4 ± 5.2
	+CFM ($\tau^{\text{fail}} = 24$ dB)	34.7 ± 1.2	9.5	5.0 ± 4.9	34.8 ± 1.2	4.5	6.9 ± 6.0
	+CFM ($\tau^{\text{fail}} = 26$ dB)	34.9 ± 1.2	9.3	8.9 ± 6.6	35.1 ± 1.2	5.1	17.7 ± 9.7
HKvasir	Bicubic	26.0 ± 0.7	–	–	26.0 ± 0.7	–	–
	BasicVSR	30.4 ± 1.2	–	–	30.4 ± 1.2	–	–
	+CFM ($\tau^{\text{fail}} = 22$ dB)	37.4 ± 2.1	10.2	5.6 ± 2.6	38.3 ± 2.1	5.0	10.0 ± 4.3
	+CFM ($\tau^{\text{fail}} = 24$ dB)	37.9 ± 2.1	10.2	7.5 ± 3.4	38.7 ± 2.0	5.0	13.7 ± 5.3
	+CFM ($\tau^{\text{fail}} = 26$ dB)	38.3 ± 2.1	10.2	10.3 ± 4.4	39.2 ± 2.0	5.0	19.4 ± 6.3

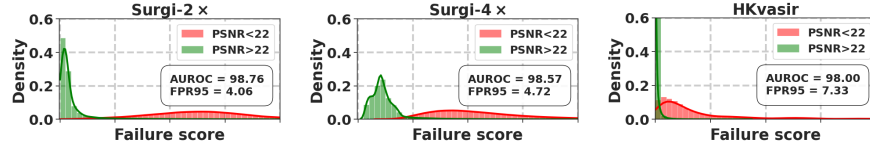


Fig. 2: Reconstruction Error Network' scores performance.

viding relatively low FNR95 results. We present the results of our constructed failure masks in Table 1. First, it is worth noting that all SR models yield improved reconstructions, including SwinIR models pre-trained on natural datasets (+2.2 and +0.7 in PSNR on the SurgiSR datasets), compared to bicubic interpolation. Second, PSNR values significantly improve when considering only the non-rejected areas after introducing the CFMs ($\geq +2.2$, $\geq +0.9$, and $\geq +7.0$ for each dataset, respectively). The empirical miscoverage rate at each failure level closely aligns with the target α value while maintaining a reasonably small average rejected area, e.g., 4.0% in Surgi-2 $\times$ ($\tau^{\text{fail}} = 22$ dB, $\alpha = 0.05$), ensuring practical utility. Note that, for challenging domains such as 4K images or the noisy HKvasir, these masks are slightly increased to satisfy the desired risk level. **Comparison with Adame *et al.* [1].** This recent effort to construct SR rejection masks with guarantees focuses on controlling the maximum error per image. In [1], the meaning of α differs: it does not control miscoverage but rather the allowed peak error; e.g., $\alpha_{[1]}$ denotes the minimum PSNR among the non-rejected pixels. In practice, this results in successful, but overly conservative masks that cover almost the entire image (see Fig. 3(a), $\alpha_{[1]} = 22$ dB). Even though $\alpha_{[1]}$

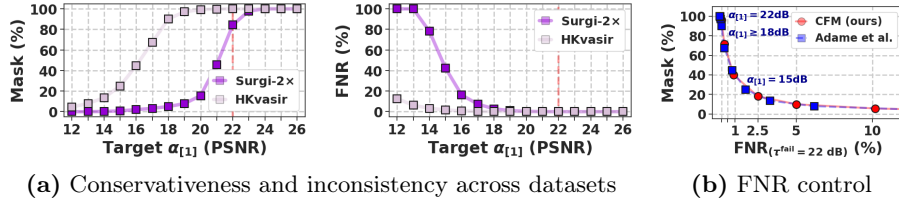
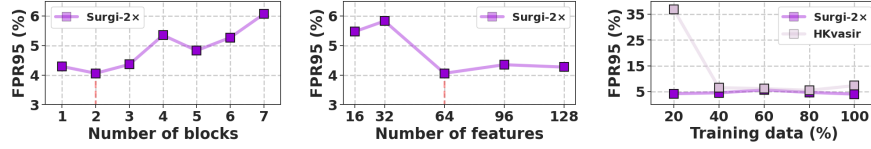
Fig. 3: Comparison of our CFM with Adame *et al.* [1] procedure.

Fig. 4: Reconstruction Error Network configuration and data-efficiency.

can be adjusted to reduce the mask size, the procedure in [1] lacks direct control over miscoverage at each desired failure level, as this trade-off is task-specific. In contrast, our CFMs simultaneously control the desired failure level and FNR, making them a more flexible solution. We aim to exemplify such a lack of control in Fig. 3(b) on the HKvasir dataset. Concretely, we plot the FNR for a target failure level of $\tau^{\text{fail}} = 22$ dB against the average mask size of different operative thresholds with α (%) = {0.1, 0.5, 1, 2.5, 5, 10} for our method (*red dots*), against $\alpha_{[1]}$ (dB) = {13, ..., 22, ..., 26} (*blue squares*) for the procedure in [1]. As previously pointed out, [1] does not produce almost any false negatives, but at the cost of creating large masks when applied at the target PSNR level, i.e., $\alpha_{[1]} = 22$ dB. For example, if one would require more relaxed masks for usability purposes, e.g., covering $\leq 20\%$ of the image on average, then one should set $\alpha_{[1]} \sim 15$ dB or below, which is quite arbitrary, and is not linked with any particular FNR level. In contrast, for our bi-level procedure, as per design in Eq. 2, FNR is successfully controlled by α at each τ^{fail} . Indeed, if more conservative masks were desired, the user would only need to decrease α , as shown in Fig. 3(b). Hence, the conformal procedure in [1] can be seen as a specific, conservative case of our general framework, where we set $\alpha \rightarrow 0$ for a given τ^{fail} .

Network configuration. Fig. 4 demonstrates effectiveness with reduced complexity (*left & center*), and promising data efficiency (*right*), requiring only a few videos for optimal performance. Note also that this module was trained on HKvasir with videos already used for validation during SR model pre-training (e.g., hyper-parameter tuning), indicating successful reuse of this data.

CFMs visualizations. Fig. 1 (*best zoomed-in*) shows the produced CFMs for Surgi-4x (*top*) and HKvasir (*bottom*). These typically detect challenging light-

ing conditions, e.g., reflections (B), incorrectly reconstructed blood vessels (A), smoke artifacts along borders (C), and overly smoothed patterns (G).

4 Discussion

This work demonstrates the potential of our CFMs to localize erroneous regions in endoscopic SR. While they guarantee control over false negatives, their quality depends on accurate error prediction, as poor estimates can yield overly conservative masks; improving these predictions under real-time constraints remains future work. Our guarantees inherit known limitations of conformal methods in medical imaging [34], including reliance on calibration size, marginal coverage, and exchangeability assumptions. Although calibration size is less restrictive at the pixel level, the latter assumptions warrant further investigation [29,17,6,46].

Acknowledgments. Innovation project supported by Innosuisse (122.308 IP-LS).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Adame, E., Csillag, D., Goedert, G.T.: Image super-resolution with guarantees via conformalized generative models. In: NeurIPS (2025) 2, 5, 6, 7, 8
2. Alex Gammerman, Volodya Vovk, V.V.: Learning by transduction. In: Conference on Uncertainty in Artificial Intelligence. pp. 148–156 (1998) 2
3. Angelopoulos, A., et al.: Uncertainty sets for image classifiers using conformal prediction. In: ICLR (2020) 2
4. Angelopoulos, A.N., et al.: Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In: ICML (2022) 2
5. Angelopoulos, A.N., et al.: Conformal risk control. In: ICLR (2024) 3, 5
6. Barber, R., et al.: Conformal prediction beyond exchangeability. *The Annals of Statistics* **51** (04 2023) 9
7. Bereska, J.I., et al.: Sacp: Spatially-adaptive conformal prediction in uncertainty quantification of medical image segmentation. In: MIDL (2025) 2
8. Borgli, H., et al.: Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific data* **7**(1), 283 (2020) 6
9. Brunekreef, J., et al.: Kandinsky Conformal Prediction: Efficient Calibration of Image Segmentation Algorithms. In: CVPR (2024) 2
10. Chan, K.C., et al.: Basicvsr: The search for essential components in video super-resolution and beyond. In: CVPR (2021) 1, 5, 6
11. Chan, K.C., et al.: Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In: CVPR (2022) 1, 5, 6
12. Chen, A., et al.: Modeling and Understanding Uncertainty in Medical Image Classification. In: MICCAI (2024) 2
13. Cheung, M.Y., et al.: Metric-guided image reconstruction bounds via conformal prediction. In: MICCAI (2025) 2

14. Chow, L.S., Paramesran, R.: Review of medical image quality assessment. *Biomedical signal processing and control* **27**, 145–154 (2016) [1](#)
15. Corbière, C., et al.: Addressing failure prediction by learning model confidence. In: *NeurIPS* (2019) [2](#), [6](#)
16. Corbière, C., et al.: Confidence estimation via auxiliary models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(10), 6043–6055 (2022) [2](#)
17. Ding, T., et al.: Class-conditional conformal prediction with many classes. In: *NeurIPS* (2023) [9](#)
18. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: *ICML* (2016) [2](#)
19. Goh, H.K.C., et al.: Minimally invasive surgery for head and neck cancer. *The lancet oncology* **11**(3), 281–286 (2010) [1](#)
20. Granese, F., et al.: Doctor: A simple method for detecting misclassification errors. In: *NeurIPS* (2021) [6](#)
21. Guo, C., et al.: On calibration of modern neural networks. In: *ICML* (2017) [2](#)
22. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: *ICLR* (2017) [2](#)
23. Jiang, F., et al.: Surgisr4k: A high-resolution endoscopic video dataset for robotic-assisted minimally invasive procedures. *Machine Learning for Biomedical Imaging* **3**, 875–885 (2025) [6](#)
24. Lafon, M., et al.: Vilu: Learning vision-language uncertainties for failure prediction. In: *ICCV* (2025) [2](#), [6](#)
25. Lakshminarayanan, B., et al.: Simple and scalable predictive uncertainty estimation using deep ensembles. *NeurIPS* **30** (2017) [2](#)
26. Lambert, B., et al.: Robust conformal volume estimation in 3d medical images. In: *MICCAI* (2024) [2](#)
27. Liang, J., et al.: Swinir: Image restoration using swin transformer. In: *ICCV Workshops* (2021) [1](#), [6](#)
28. Liu, X., et al.: Medvsr: Medical video super-resolution with cross state-space propagation. In: *ICCV* (2025) [1](#), [5](#), [6](#)
29. Löfström, T., et al.: Bias reduction through conditional conformal prediction. *Intelligent Data Analysis* **19**(6), 1355–1375 (2015) [9](#)
30. Lu, C., et al.: Fair conformal predictors for applications in medical imaging. In: *AAAI* (2022) [2](#)
31. Lu, C., et al.: Improving trustworthiness of ai disease severity rating in medical imaging with ordinal conformal prediction sets. In: *MICCAI* (2022) [2](#)
32. Mack, M.J.: Minimally invasive and robotic surgery. *Jama* **285**(5), 568–572 (2001) [1](#)
33. Mauricio Sadinle, J.L., Wasserman, L.: Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association* **114**(525), 223–234 (2019) [2](#)
34. Mehrtens, H.A., et al.: Pitfalls of conformal predictions for medical image classification. In: *UNSURE, MICCAI Workshops* (2023) [9](#)
35. Mossina, L., Friedrich, C.: Conformal prediction for image segmentation using morphological prediction sets. In: *MICCAI* (2025) [2](#)
36. Niu, B., et al.: Single image super-resolution via a holistic attention network. In: *ECCV* (2020) [1](#)
37. Papadopoulos, H., et al.: Inductive confidence machines for regression. In: *ECML*. pp. 345–356 (2002) [2](#)
38. Romano, Y., et al.: Conformalized quantile regression. In: *NeurIPS* (2019) [2](#)

- 39. Sangalli, S., et al.: Conformal forecasting for surgical instrument trajectory. In: MICCAI (2025) 2
- 40. Si, W., et al.: Reliable and interpretable visual field progression prediction with diffusion models and conformal risk control. In: MICCAI (2025) 2
- 41. Silva-Rodríguez, J., et al.: Full conformal adaptation of medical vision-language models. In: IPMI (2025) 2
- 42. Silva-Rodríguez, J., et al.: Trustworthy few-shot transfer of medical vlms through split conformal prediction. In: MICCAI (2025) 2
- 43. Stankeviciute, K., et al.: Conformal time-series forecasting. In: NeurIPS (2021) 2
- 44. Teneggi, J., et al.: How to trust your diffusion model: A convex optimization approach to conformal risk control. In: ICML (2023) 2
- 45. Teneggi, J., et al.: Conformal risk control for semantic uncertainty quantification in computed tomography. In: MICCAI (2025) 2
- 46. Tibshirani, R.J., et al.: Conformal prediction under covariate shift. In: NeurIPS (2019) 9
- 47. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer (2005) 2