A Bayesian Approach to Segmentation with Noisy Labels via Spatially Correlated Distributions

Anonymous authors
Paper under double-blind review

Abstract

In semantic segmentation, the accuracy of models heavily depends on the high-quality annotations. However, in many practical scenarios, such as medical imaging and remote sensing, obtaining true annotations is not straightforward and usually requires significant human labor. Relying on human labor often introduces annotation errors, including mislabeling, omissions, and inconsistency between annotators. In the case of remote sensing, differences in procurement time can lead to misaligned ground-truth annotations. These label errors are not independently distributed, and instead usually appear in spatially connected regions where adjacent pixels are more likely to share the same errors. To address these issues, we propose an approximate Bayesian estimation based on a probabilistic model that assumes training data include label errors, incorporating the tendency for these errors to occur with spatial correlations between adjacent pixels. However, Bayesian inference for such spatially correlated discrete variables is notoriously intractable. To overcome this fundamental challenge, we introduce a novel class of probabilistic models, which we term the **ELBO-Computable Correlated Discrete Distribution (ECCD)**. By representing the discrete dependencies through a continuous latent Gaussian field with a Kac-Murdock-Szegö (KMS) structured covariance, our framework enables scalable and efficient variational inference for problems previously considered computationally prohibitive. Through experiments on multiple segmentation tasks, we confirm that leveraging the spatial correlation of label errors significantly improves performance. Notably, in specific tasks such as lung segmentation, the proposed method achieves performance comparable to training with clean labels under moderate noise levels. Code is included in the supplementary materials.

1 Introduction

Semantic segmentation, which involves classifying each pixel in an image into one of several classes, is a crucial task in computer vision. In supervised learning, the accuracy of segmentation models critically depends on the quality of the annotations in the training data. However, obtaining truly accurate pixel-level annotations is challenging in many practical applications. Even when expert annotators are employed, errors, omissions, and subjectivity in interpretation are inevitable, leading to inconsistencies in the datasets. In particular, high inter- and intra-annotator variability is widely reported in medical imaging, where experts may have differing interpretations of the same structures. For instance, several studies Zhang et al. (2020a); Lampert et al. (2016); Rädsch et al. (2023); Yang et al. (2023) highlight significant discrepancies among expert annotators; some delineate structures more generously, while others prefer more conservative annotations. Observer-dependent annotations exacerbate label noise in supervised learning. Label noise is also a critical issue in remote sensing, where determining ground truth labels often requires field surveys over large and sometimes inaccessible regions Frénay & Verleysen (2013); Pelletier et al. (2017); Foody (2002). Due to the logistical and economic challenges of large-scale ground truth collection, researchers frequently rely on automatic labeling systems, which may introduce systematic errors. Additionally, high-quality annotated datasets remain a critical bottleneck for supervised learning, particularly in remote sensing applications where annotations are often repurposed across different types of satellite images. For example, the OpenEarthMap dataset Xia et al. (2023) was created by manually annotating high-resolution optical satellite images for

semantic segmentation. However, these annotations are sometimes reused for synthetic aperture radar (SAR) imagery, despite differences in resolution and capture conditions Huang et al. (2021); Zhang et al. (2021); Liu et al. (2023). Additionally, changes in artificial structures or variations in land cover further contribute to label inconsistencies Fritz et al. (2009); Huang et al. (2021).

Various approaches have been proposed to mitigate the adverse effects of noisy labels. Some methods attempt to stop training early to prevent the network from overfitting to noise and generating unreliable pseudo-labels Liu et al. (2024; 2022), while others modify the loss function to be more robust against large errors Gonzalez-Jimenez et al. (2025). Although these techniques can reduce the influence of noisy annotations, they do not fundamentally address the core reason why the standard supervised learning framework fails in the presence of noisy labels.

We propose a method that directly tackles the root cause of this issue. To understand this root cause, we first revisit the probabilistic assumptions underlying the standard cross-entropy loss. In supervised learning for segmentation models, training typically reduces to optimizing this loss function. This optimization implicitly follows a maximum likelihood estimation (MLE) framework under the assumption that the training data consists of independent and identically distributed samples drawn from the joint distribution of images and clean labels. However, when labels are noisy, the assumption of identically distributed assumption no longer holds, as the observed labels systematically deviate from clean labels, leading to performance degradation in the trained model.

To address this issue, we maintain the MLE framework but generalize the underlying probabilistic model to make it more suitable for real-world scenarios with noisy labels. Specifically, we introduce a model that explicitly accounts for the presence of noisy labels, which differ from the clean labels due to labeling errors. In practice, annotation errors tend to exhibit strong spatial correlations - mislabeling often occurs in contiguous regions. Variations in annotation criteria among experts, as well as changes in the underlying scene — such as the construction or demolition of buildings or alterations in vegetation — further reinforce this spatial dependency. Given this, we assume that label errors are not independent but instead exhibit spatial correlations among pixels.

However, directly modeling spatial correlations among discrete labels leads to an intractable marginal likelihood, a long-standing challenge in probabilistic modeling Koller & Friedman (2009); Sutton & McCallum (2012). Our core contribution is to resolve this challenge by proposing a novel class of tractable models, the **ELBO-Computable Correlated Discrete Distribution (ECCD)**. This discrete distribution is represented through continuous variables that follow a Gaussian distribution, where spatial correlations between pixels are expressed via a covariance matrix. Representing discrete variables through a Gaussian distribution successfully circumvents the intractability of summing over all possible realizations of the discrete variables. While the covariance matrix, whose number of elements scales quadratically with the number of pixels, introduces additional computational challenges in evaluating the ELBO, particularly in computing second-order statistics, its inverse, and its determinant. To overcome these computational intractabilities, we leverage the Kac-Murdock-Szegö (KMS) matrix Fikioris (2018); Kac et al. (1953), which enables efficient computations necessary for ELBO evaluation. To validate the effectiveness of our approach, we conduct extensive empirical evaluations on multiple segmentation tasks. Our experimental results demonstrate that our method significantly improves robustness against label noise, particularly in scenarios with moderate to high levels of spatially correlated label noise. In summary, our contributions are as follows:

1. **A Novel Class of Tractable Distributions for Correlated Discrete Variables:**
   We introduce the **ECCD**, a new class of probabilistic models for tractable inference on high-dimensional discrete variables with stationary, exponentially decaying correlations—a structure common to spatial grids and temporal sequences. In this work, we apply the ECCD to learn from spatially correlated label noise in medical and remote sensing, where the model circumvents computational intractability by representing discrete dependencies via a continuous latent Gaussian field.

2. **Efficient ELBO optimization via the KMS matrix:**
   The ECCD's covariance matrix introduces computational challenges that scale quadratically with the number of pixels. We overcome this by leveraging the Kac-Murdock-Szegö (KMS) matrix. To the best of our knowledge, this is the first work to introduce the KMS matrix as a scalable computational

tool for variational inference with correlated latent variables, enabling efficient computation of the key operations necessary for optimizing the ELBO of the ECCD.

3. **Extensive empirical validation:**
We perform comprehensive experiments on multiple segmentation tasks, demonstrating that our method significantly improves robustness to label noise. Our approach outperforms existing techniques, particularly in scenarios with moderate to high levels of spatially correlated label noise, validating the effectiveness of our probabilistic formulation.

2 Related Work

Learning with noisy labels has been a significant focus in the field of machine learning, particularly classification tasks Patrini et al. (2017); Yao et al. (2020); Han et al. (2018); Wei et al. (2020). One research direction models the relationship between the clean label y^* and the noisy label y through a transition matrix Patrini et al. (2017); Yao et al. (2020) that characterizes the probabilities from the clean label to the noisy label. For example, the transition probability from a clean label of k to a noisy label of j is represented as $p(y = j|y^* = k)$. Additionally, methods have been proposed to estimate an instance-dependent transition matrix Xia et al. (2020); Berthon et al. (2021), where the transition probabilities depend on the instance x also, denoted as $p(y = j|y^* = k, x)$. These classification methods can be adapted for semantic segmentation tasks; however, they treat label errors as being independent to each pixel, thereby ignoring the spatial correlations in label errors that may be present in annotations. This limitation is particularly critical in applications such as medical imaging Zhang et al. (2020a); Shi & Wu (2021); Zhang et al. (2020b) and remote sensing Mnih & Hinton (2012); Song et al. (2022); Liu et al. (2024); Henry et al. (2021), where different types of noise may occur.

Other lines of research includes work by Zhang *et al.* Zhang et al. (2020c), which proposed a Tri-Network framework that trains three networks simultaneously, with each pair selecting reliable pixels based on their loss maps, thereby achieving robust learning even with coarse annotations. Liu *et al.* Liu et al. (2022) introduced a mechanism to detect the moment when a network begins to overfit noisy labels. Gonzalez *et al.* Gonzalez-Jimenez et al. (2025) designed T-Loss based on a Student-t distribution to apply a logarithmic penalty on large errors, reducing undue influence from outlier pixels. Nonetheless, these methods do not take into account the spatial correlations in label errors that may occur within annotations.

Unlike the aforementioned related works, there are few studies that consider spatial correlations. Li *et al.* Li et al. (2021) leveraged superpixel segmentation to group pixels and smooth the network outputs within each superpixel, thereby indirectly incorporating spatial context. This method, however, relies heavily on the quality of the generated superpixels. More recently, Yao *et al.* Yao et al. (2023) proposed an approach that explicitly models spatial correlations by applying a Markov process that consists of expansion and shrinkage steps of the annotation masks. Although this method achieves better performance, it is mainly limited to addressing noise along the boundaries. In real-world scenarios, segmentation labels are often corrupted by a diverse range of noise types, including omission noise, misalignment noise, boundary uncertainties, and other systematic labeling errors Jiménez & Decaestecker (2024); Vădineanu et al. (2022); Liu et al. (2024); Mnih & Hinton (2012); Jiang et al. (2021).

Our proposed probabilistic model can incorporate spatial correlations in label errors without making assumptions about the noise types, thereby handling a wide range of noise types.

M

3 Problem Formulation

3.1 Probabilistic Foundations of Standard Supervised Segmentation

In standard supervised learning for semantic segmentation, the training process is typically framed as the minimization of the cross-entropy loss. This objective function is not arbitrary; it is derived from the principle of Maximum Likelihood Estimation (MLE) under a specific, and often implicit, probabilistic model. This

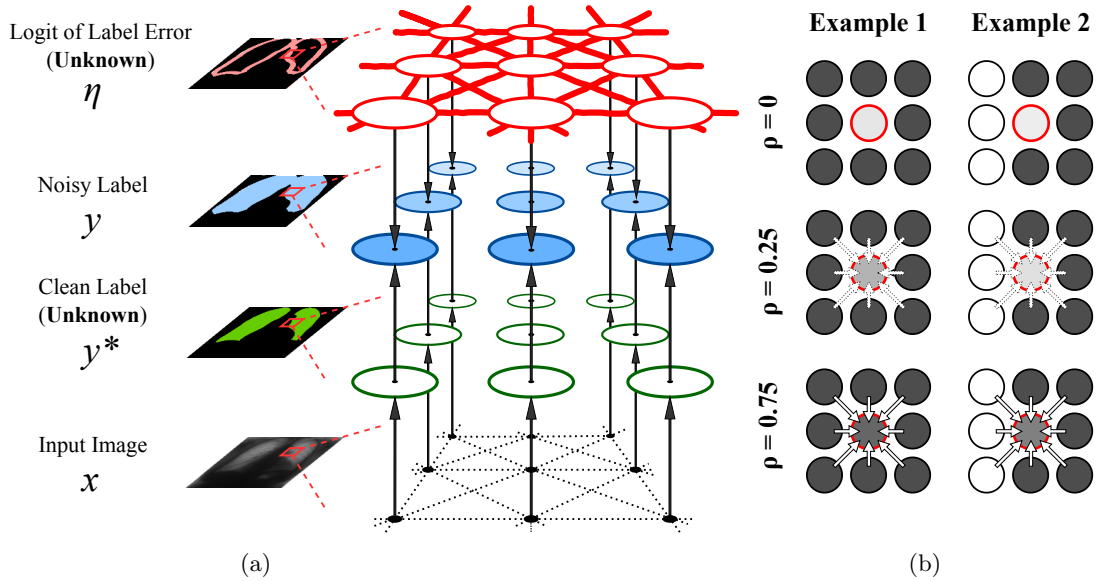


Figure 1: **(a) Graphical model of our proposed method, the ECCD.** The model introduces a latent Gaussian variable $\boldsymbol{\eta}$ that encodes spatial correlations of label error among adjacent pixels, thereby enabling a more realistic modeling of the dependency between pixels.
(b) Conceptual illustration of spatial correlation. Examples 1 and 2 show how varying ρ affects the correlation strength between the center pixel and its neighbors (top, bottom, sides, and diagonals). When adjacent pixels are likely to contain label errors and spatial correlation is strong (ρ is high) label errors can be inferred with high confidence, even if they are rare.

underlying probabilistic model makes two fundamental assumptions: (1) the provided annotations $\mathbf{y}$ are the true, noise-free clean labels $\mathbf{y}^*$, and (2) the label of each pixel y_i is conditionally independent of the labels of other pixels given the input image $\mathbf{x}$.

Under these assumptions, the goal is to find the model parameters θ that maximizes the log-likelihood of the observed data:

$$\log p_\theta(\mathbf{y}|\mathbf{x}) = \log \prod_{i=1}^{HW} p_\theta(y_i|\mathbf{x}) = \sum_{i=1}^{HW} \log p_\theta(y_i|\mathbf{x}), \tag{1}$$

where HW is the total number of pixels. The segmentation model $p_\theta(y_i|\mathbf{x})$ outputs a probability distribution over the set of possible classes $\mathcal{Y}$ for each pixel i. Let this distribution be represented by a vector of probabilities $\{p_\theta(y_{ik}|\mathbf{x})\}_{k=1}^{|\mathcal{Y}|}$. If we represent the ground-truth label y_i as a one-hot vector, where $y_{ik} = 1$ if pixel i belongs to class k and $y_{ik} = 0$ otherwise, the log-likelihood for a single pixel can be written as $\sum_{k=1}^{|\mathcal{Y}|} y_{ik} \log p_\theta(y_{ik}|\mathbf{x})$. The total log-likelihood for the image is therefore:

$$\mathcal{L}_{\text{MLE}}(\theta) = \sum_{i=1}^{HW} \sum_{k=1}^{|\mathcal{Y}|} y_{ik} \log p_\theta(y_{ik}|\mathbf{x}). \tag{2}$$

Minimizing the negative of this quantity, $-\mathcal{L}_{\text{MLE}}(\theta)$, is precisely equivalent to minimizing the standard multi-class cross-entropy loss. This derivation highlights that the validity of the cross-entropy loss is contingent on the assumption that the training labels are clean and pixel-wise independent samples from the true data distribution.

3.2 A General Probabilistic Model for Learning with Noisy Labels

In many practical scenarios, particularly in medical imaging and remote sensing, the assumption that observed labels $\mathbf{y}$ are identical to the true clean labels $\mathbf{y}^*$ does not hold. Annotations are often corrupted by noise, which can be spatially correlated. To address this, we generalize the probabilistic model by explicitly distinguishing between the observed labels $\mathbf{y}$ and the unobserved (latent) clean labels $\mathbf{y}^*$. This leads to the objective of maximizing the marginal log-likelihood, which integrates out the uncertainty over the unknown clean labels $\mathbf{y}^*$:

$$\log p_\theta(\mathbf{y}|\mathbf{x}) = \log \left(\sum_{\mathbf{y}^*} p_\theta(\mathbf{y}, \mathbf{y}^*|\mathbf{x}) \right) = \log \left(\sum_{\mathbf{y}^*} p(\mathbf{y}|\mathbf{y}^*) p_\theta(\mathbf{y}^*|\mathbf{x}) \right). \tag{3}$$

Here, $p_\theta(\mathbf{y}^*|\mathbf{x})$ is the probability of the clean labels $\mathbf{y}^*$ given the image $\mathbf{x}$, which is modeled by the segmentation network, and $p(\mathbf{y}|\mathbf{y}^*)$ represents the label noise process.

This formulation is a direct generalization of the standard MLE objective described in Section 3.1. To see this, consider the noise-free case where the noise model $p(\mathbf{y}|\mathbf{y}^*)$ collapses to a Kronecker delta function $\delta(\mathbf{y}^*, \mathbf{y})$, which equals 1 if and only if $\mathbf{y}^* = \mathbf{y}$, and is zero otherwise. In this scenario, the summation in Eq. (3) contains only one non-zero term, and the objective reduces precisely to maximizing $\log p_\theta(\mathbf{y}|\mathbf{x})$, which is the basis for the standard cross-entropy loss as shown in Eq. (1). However, the summation over all possible configurations of $\mathbf{y}^*$ is computationally intractable, as the number of terms ($|\mathcal{Y}|^{HW}$) grows exponentially with the number of pixels. One way to overcome this is to assume a pixel-independent noise process, $p(\mathbf{y}|\mathbf{y}^*) = \prod_i p(y_i|y_i^*)$, but this fails to capture real-world spatially correlated label noise, as mentioned in Section 1.

4 The Proposed ECCD Framework

4.1 A Tractable Variational Objective via the ECCD

To overcome the intractability of the marginal likelihood while modeling spatial correlations, we resort to variational inference and maximize its Evidence Lower Bound (ELBO). We introduce a continuous latent variable, $\boldsymbol{\eta} \in \mathbb{R}^{HW}$, which represents the logit of the label error probability for each pixel. This variable captures spatial correlations and indirectly imposes spatial dependencies on the labels as follows:

$$p(\mathbf{y}|\mathbf{y}^*) = \int p(\mathbf{y}|\mathbf{y}^*, \boldsymbol{\eta}) p(\boldsymbol{\eta}) d\boldsymbol{\eta} = \int \left\{ \prod_i p(y_i|y_i^*, \eta_i) \right\} p(\boldsymbol{\eta}) d\boldsymbol{\eta}. \tag{4}$$

Figure 1a depicts the graphical model illustrating this assumption. To fully specify this generative model, which is the central component of the ECCD, we define its constituent distributions. The prior $p(\boldsymbol{\eta})$ is modeled as a Gaussian, $p(\boldsymbol{\eta}) = \mathcal{N}(\boldsymbol{\eta}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with its covariance matrix structured using a Kac-Murdock-Szegö (KMS) matrix Fikioris (2018); Kac et al. (1953), which we detail in Section 4.2. The conditional noise model $p(y_i|y_i^*, \eta_i)$ is defined as:

$$p(y_{ik} = 1|y_{ic}^* = 1, \eta_i) = \delta_{kc}(1 - r(\eta_i)) + (1 - \delta_{kc}) r(\eta_i) W_{kc}, \tag{5}$$

where δ_{kc} is the Kronecker delta, $r(\cdot)$ is the sigmoid function that maps the latent variable η_i to a label error probability, and W is a learnable class transition matrix that specifies the probability distribution over incorrect labels when an error occurs, with its elements satisfying $W \in [0,1]^{|\mathcal{Y}| \times |\mathcal{Y}|}$, $W_{cc} = 0$ and $\sum_{k=1}^{|\mathcal{Y}|} W_{kc} = 1$, for each true class $c \in \mathcal{Y}$.

To derive a tractable objective for this model, we introduce a variational distribution $q(\mathbf{y}^*, \boldsymbol{\eta}|\mathbf{y})$ to approximate the true posterior. We assume a factorized form symmetric to the generative model: $q(\mathbf{y}^*, \boldsymbol{\eta}|\mathbf{y}) = q(\boldsymbol{\eta}) \prod_i q(y_i^*|y_i, \eta_i)$. The variational posterior for $\boldsymbol{\eta}$ is also a Gaussian, $q(\boldsymbol{\eta}) = \mathcal{N}(\boldsymbol{\eta}|\boldsymbol{m}, \boldsymbol{\Gamma})$, and the conditional posterior for the clean labels is parameterized as:

$$q(y_{ik}^* = 1|y_{ic} = 1, \eta_i) = \delta_{kc}(1 - r(\eta_i)) + (1 - \delta_{kc}) r(\eta_i) V_{kc}, \tag{6}$$

where $V \in [0,1]^{|\mathcal{Y}| \times |\mathcal{Y}|}$ is another learnable class transition matrix modeling the inverse noise process (i.e., inferring the clean label from the noisy one), with analogous constraints ($V_{cc} = 0$ and $\sum_{k=1}^{|\mathcal{Y}|} V_{kc} = 1$ for each

noisy class $c \in \mathcal{Y}$). With the generative model and variational posterior fully defined, we can derive the Evidence Lower Bound (ELBO) as follows (see Appendix A.2 for details):

$$\begin{aligned}\log p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta}) &\geq \mathbb{E}_{q(\mathbf{y}^*,\boldsymbol{\eta}|\mathbf{y})}\left[\log p(\mathbf{y}|\mathbf{y}^*,\boldsymbol{\eta},\mathbf{x},\boldsymbol{\theta})\right] - KL\left[q(\mathbf{y}^*,\boldsymbol{\eta}|\mathbf{y}) \,\|\, p(\boldsymbol{\eta},\mathbf{y}^*|\mathbf{x},\boldsymbol{\theta})\right] \\ &= \sum_i \mathbb{E}_{q(\eta_i)q(y_i^*|y_i,\eta_i)}\left[\log p_{\boldsymbol{\theta}}(y_i^*|\mathbf{x})\right] - \mathbb{E}_{q(\boldsymbol{\eta})}\left[\log q(\boldsymbol{\eta})\right] + \mathbb{E}_{q(\boldsymbol{\eta})}\left[\log p(\boldsymbol{\eta})\right] \\ &\quad - \sum_i \mathbb{E}_{q(\eta_i)q(y_i^*|y_i,\eta_i)}\left[\log q(y_i^*|y_i,\eta_i)\right] + \sum_i \mathbb{E}_{q(\eta_i)q(y_i^*|y_i,\eta_i)}\left[\log p(y_i|y_i^*,\eta_i)\right]. \end{aligned} \tag{7}$$

A key advantage of this formulation is that the expectations over the discrete latent labels $\mathbf{y}^*$ in Eq.(7) are now decomposed into a sum over individual pixels. This avoids the intractable summation over all possible label map configurations that hindered the direct optimization of the marginal likelihood in Eq.(3), rendering the objective tractable. The overall goal is to optimize the model parameters $\boldsymbol{\theta}$ and the variational parameters of $q(\boldsymbol{\eta})$ and $q(y_i^*|y_i,\eta_i)$ to maximize this lower bound. The ELBO in Eq.(7) has an important interpretation as a principled generalization of the standard cross-entropy loss. To see this, we can analyze its components.

The first term in this ELBO (7) can be interpreted as an expected reconstruction term. By expanding the expectation over the discrete clean label y_i^*, it becomes a cross-entropy loss with soft labels:

$$\sum_i \mathbb{E}_{q(y_i^*|y_i,\eta_i)q(\eta_i)}\left[\log p_{\boldsymbol{\theta}}(y_i^*|\mathbf{x})\right] = \sum_{i=1}^{HW}\sum_{k=1}^{|\mathcal{Y}|} q(y_{ik}^*=1|y_i)\log p_\theta(y_{ik}^*=1|\mathbf{x}), \tag{8}$$

where $q(y_{ik}^*=1|y_i) = \mathbb{E}_{q(\eta_i)}[q(y_{ik}^*=1|y_i,\eta_i)]$ is the marginal posterior probability that the clean label for pixel i is class k. This replaces the hard, fixed one-hot observed labels with soft, inferred posterior probabilities. Building on this interpretation of the first term, we can now demonstrate more formally that our entire objective is a principled generalization of the standard cross-entropy loss. This is shown by examining the conditions under which the remaining terms in the ELBO vanish.

The second and third terms together constitute the negative Kullback-Leibler (KL) divergence, $-\mathrm{KL}[q(\boldsymbol{\eta})||p(\boldsymbol{\eta})]$. This term vanishes when the variational posterior is identical to the prior, i.e., $q(\boldsymbol{\eta}) = p(\boldsymbol{\eta})$. Similarly, the fourth and fifth terms represent the expected log-likelihood of the observed labels under the noise model, regularized by the entropy of the label posterior. These terms also cancel each other out under the assumption of a symmetric noise process, where $V = W$ in Eqs. (6) and (5). This condition is met in non-trivial noisy scenarios, for instance, with a uniform noise model where an error from any class is equally likely to transition to any other incorrect class (i.e., $W_{kc} = V_{kc} = 1/(|\mathcal{Y}|-1)$ for $k \neq c$).

In such a setting, maximizing the ELBO is equivalent to maximizing only the first term. As we saw, if we assume a noise-free setting, the posterior distribution $q(y_i^*|y_i,\eta_i)$ collapses to a point mass (i.e., a Kronecker delta function centered at the observed label y_i), making the soft labels $q(y_{ik}^*|y_i)$ identical to the observed one-hot labels. Consequently, our objective function reduces precisely to the standard cross-entropy loss. Therefore, our method provides a principled generalization of the standard supervised learning framework, where the additional terms in the ELBO serve to regularize the posterior and enforce spatial correlation in the inferred label errors.

While the expectations over the discrete labels y_i^* are tractable, the terms involving the continuous variable $\boldsymbol{\eta}$ pose a significant computational challenge. The number of elements in $\boldsymbol{\eta}$ equals the number of image pixels (HW), which can be in the hundreds of thousands, and these elements are spatially correlated. For instance, the negative KL divergence term, $-\mathrm{KL}[q(\boldsymbol{\eta})||p(\boldsymbol{\eta})]$, is analytically computable since both $p(\boldsymbol{\eta}) = \mathcal{N}(\boldsymbol{\eta}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ and $q(\boldsymbol{\eta}) = \mathcal{N}(\boldsymbol{\eta}|\boldsymbol{m},\boldsymbol{\Gamma})$ are Gaussians. Its value is given by:

$$-\mathrm{KL}[q(\boldsymbol{\eta})||p(\boldsymbol{\eta})] = \frac{1}{2}\left(\log\frac{|\boldsymbol{\Gamma}|}{|\boldsymbol{\Sigma}|} + HW - \mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma}\right) - (\boldsymbol{m}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{m}-\boldsymbol{\mu})\right). \tag{9}$$

However, evaluating this term is computationally prohibitive for large images, as it requires computing the determinants and inverses of the $HW \times HW$ covariance matrices, in addition to the trace and quadratic form terms. These challenges are efficiently addressed by leveraging the properties of the KMS-structured covariance, as we describe next.

4.2 KMS-structured covariance

An $n \times n$ KMS matrix R_ρ is a type of symmetric Toeplitz matrix, where each element is defined using the parameter $\rho \in (-1, 1)$. Its inverse, R_ρ^{-1}, is a tridiagonal matrix. The explicit forms of R_ρ and R_ρ^{-1} are given by:

$$R_\rho := \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{pmatrix}, \quad R_\rho^{-1} = \tfrac{1}{1-\rho^2} \begin{pmatrix} 1 & -\rho & 0 & \cdots & 0 \\ -\rho & 1+\rho^2 & -\rho & \cdots & 0 \\ 0 & -\rho & 1+\rho^2 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & -\rho \\ 0 & 0 & \cdots & -\rho & 1 \end{pmatrix},$$

and the determinant is computed as $|R_\rho| = (1 - \rho^2)^{n-1}$. See Appendix A.1 for mathematical derivations. While the KMS matrix is well-established in fields like signal processing for modeling AR(1) processes Grenander & Szegö (1958); Brockwell & Davis (1991), its unique properties — a sparse inverse and an analytic determinant — have been largely untapped as a computational tool in modern machine learning. Our work bridges this gap by demonstrating its utility as the key component for making the ECCD framework computationally tractable.

This matrix is useful for representing correlations among one-dimensional variables, where the correlation decays exponentially as the index distance between two variables increases regardless of their absolute positions.

Suppose a one-dimensional random variable $\mathbf{x}$ has a covariance structure given by

$$\mathbb{E}[(x_i - \mu_i)(x_j - \mu_j)] = \sigma_i \sigma_j \rho^{|i-j|}, \tag{10}$$

where μ_i is the mean of x_i. By utilizing the KMS matrix R_ρ, the covariance matrix can be expressed as

$$\mathbf{\Sigma} = V R_\rho V, \quad \mathbf{\Sigma}^{-1} = V^{-1} R_\rho^{-1} V^{-1}, \tag{11}$$

where V is a diagonal matrix whose (i, i)-th element is σ_i.

In the case of two-dimensional spatial correlations, such as those in images, we utilize the Kronecker product. Suppose a two-dimensional random variable $\mathbf{x}$ has a covariance structure given by

$$\mathbb{E}[(x_{ij} - \mu_{ij})(x_{uv} - \mu_{uv})] = \sigma_{ij} \sigma_{uv} \rho^{|i-u|+|j-v|}, \tag{12}$$

where μ_{ij} is the mean of x_{ij}. Its matrix form can be expressed as

$$\mathbf{\Sigma} = V(R_\rho \otimes R'_\rho)V, \quad \mathbf{\Sigma}^{-1} = V^{-1}(R_\rho^{-1} \otimes R'^{-1}_\rho)V^{-1}, \tag{13}$$

where we assume that the two-dimensional pixel index (i, j) is mapped to a one-dimensional index n. V is a diagonal matrix whose (n, n)-th element corresponds to σ_{ij}, and $\otimes$ denotes the Kronecker product.

Since V is diagonal and R_ρ^{-1} is tridiagonal, the inverse covariance $\mathbf{\Sigma}^{-1}$ is a sparse matrix. The properties of the KMS matrix provide an elegant solution to the computational challenges outlined in Section 4.1. First, the determinant of the covariance matrix can be computed analytically, bypassing the need for expensive numerical methods. Second, the sparse structure of the inverse matrix is crucial for the remaining terms. It reduces the complexity of computing the quadratic form and the trace term in the KL divergence (9) from $O((HW)^2)$ to $O(HW)$, which is linear in the number of pixels.

Beyond this computational speed-up, the sparse structure of the inverse also ensures that the quadratic form

$$(\boldsymbol{\eta} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1} (\boldsymbol{\eta} - \boldsymbol{\mu}) \tag{14}$$

involves only pairs of adjacent pixel variables. This, in turn, makes the conditional distribution $p(\eta_i \mid \boldsymbol{\eta}_{\setminus i})$ depend only on the neighboring pixels $\boldsymbol{\eta}_{\mathcal{N}_i}$, where $\boldsymbol{\eta}_{\setminus i}$ denotes all variables in $\boldsymbol{\eta}$ except η_i, and $\mathcal{N}_i$ represents the set of adjacent pixels of i.

In other words, the KMS-structured covariance ensures that the Gaussian distribution retains a Markov Random Field property while maintaining the tractability of both the inverse and the determinant of the covariance matrix. The effect of the correlation parameter ρ is illustrated in Figure 1b. As shown, the model can confidently infer label errors when adjacent pixels are also likely to contain errors and the spatial correlation is strong (i.e., ρ is high).

4.3 Optimization

We formulate the optimization problem as follows. Let N be the number of training samples. We denote the parameters of the prior distribution as $\boldsymbol{\omega}$, the image-specific parameters of the variational posterior as $\{\boldsymbol{\nu}^{(n)}\}_{n=1}^{N}$, and the negative ELBO for the n-th sample (from Eq.(7)) as $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\omega}, \boldsymbol{\nu}^{(n)})$.

Our training procedure, outlined in Algorithm 1, employs an alternating optimization scheme. For clarity, the algorithm is presented for a single sample per step, though this can be generalized to mini-batches. In each step for a given sample, the model parameters θ are updated K times, while the corresponding variational posterior parameters $\boldsymbol{\nu}^{(n)}$ are updated M times. After training, the optimized model parameters θ are used for inference on new images. Note that in Algorithm 1, the prior parameters $\boldsymbol{\omega}$ are treated as fixed,but they could also be optimized jointly with a sufficiently large training dataset.

Algorithm 1 Training loop for alternating optimization of model and posterior parameters.

$\boldsymbol{\theta}, \boldsymbol{\omega}, \{\boldsymbol{\nu}_0^{(\ell)}\}_{\ell=1}^{N} \leftarrow$ Initialize parameters
repeat
  $n \leftarrow$ Randomly sample an index from $\{1, \ldots, N\}$
  for $k = 1, \ldots, K$ **do**
    Compute $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\omega}, \boldsymbol{\nu}_0^{(n)})$ and its gradient w.r.t. $\boldsymbol{\theta}$
    $\boldsymbol{\theta} \leftarrow$ Update the model parameters using the gradient
  end for
  for $m = 0, 1, \ldots, M-1$ **do**
    Compute $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\omega}, \boldsymbol{\nu}_m^{(n)})$ and its gradient w.r.t. $\boldsymbol{\nu}_m^{(n)}$
    $\boldsymbol{\nu}_{m+1}^{(n)} \leftarrow$ Update the posterior parameters using the gradient
  end for
  $\boldsymbol{\nu}_0^{(n)} \leftarrow \boldsymbol{\nu}_M^{(n)}$
until convergence of parameters $\boldsymbol{\theta}$

5 Experiments

5.1 Dataset

Research on segmentation with noisy labels has been especially active in both medical imaging and satellite remote sensing. In this work, we focus on two widely used benchmark datasets that represent these domains: the JSRT dataset for medical imaging and the WHU Building dataset for satellite imagery. Detailed descriptions of each dataset are provided below.

It is common practice in noisy-label segmentation research to artificially introduce label noise for evaluation because most publicly available datasets come with clean annotations Li et al. (2021); Song et al. (2022); Yao et al. (2023). Following this convention, we also employ synthetic noise in our experiments, with details of our noise generation process described in Section 5.2.1.

JSRT Dataset Shiraishi et al. (2000) is a publicly available dataset provided by the Japanese Society of Radiological Technology (JSRT), which comprises chest radiographs annotated with segmentation labels for the lung fields, heart, and clavicles. All images are of size 256×256 pixels, with 199 images designated for training and 50 for testing. Following the assumption common in noisy-label segmentation research—that no clean validation data is available—we evaluated our approach using the published split of 199 training

images and 50 test images. Following a previous study Li et al. (2021), the Clavicle region was cropped to a fixed 96×224 area for training and inference.

WHU Building Dataset Ji et al. (2018) is a publicly available dataset for building segmentation containing satellite images of Christchurch with a resolution of 0.075 m covering an area of 450 km^2. Manual annotations were provided for 22,000 buildings. The satellite images were downsampled to a resolution of 0.3 m and then divided into 512×512 tiles. Among the non-overlapping tiles, 4,736 were used for training, 1,036 for validation, and 2,416 for testing. Similar to JSRT, we adopted the published split for the training and testing data.

5.2 Experimental Setup

5.2.1 Noisy Label Settings.

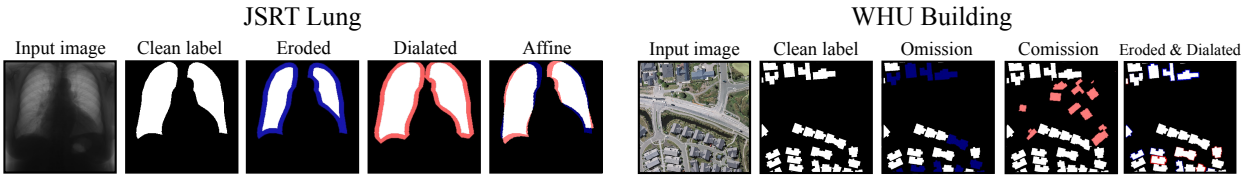


Figure 2: **Examples of Noisy Labels.** In medical imaging, label noise is often simulated using morphological operations such as erosion, dilation, and affine transformations Li et al. (2021). Conversely, building segmentation commonly involves omission noise, commission errors, and boundary noise Song et al. (2022). The pink region indicates labels present in the noisy but absent in the clean label, while the blue region indicates the opposite.

In the JSRT dataset, we evaluated the impact of noise caused by boundary error, which is a critical issue in the medical imaging domain Yao et al. (2023); Zhang et al. (2020b). In accordance with previous works Xue et al. (2020); Zhang et al. (2020c); Xue et al. (2020); Li et al. (2021), we employed morphological transformations as synthetic noise. As illustrated in Figure 2, the noise settings consider not only boundary ambiguities (i.e., dilation or erosion) but also affine transformations. We examined the effect of noise using α to represent the proportion of affected data and β to determine its strength, setting $\{(\alpha, \beta)\} = \{(0.3, 0.5), (0.5, 0.7), (0.7, 0.7)\}$.

In the WHU Building dataset, we modeled three types of label noise: omission noise, where labels are missing in regions that should be labeled Song et al. (2022); Liu et al. (2024); commission error, where labels are present in regions that should not be due to temporal changes Song et al. (2022); and boundary noise, which commonly occurs in segmentation tasks Song et al. (2022), as illustrated in Figure 2. We define the omission ratio (ϕ), the commission ratio (ζ), and the boundary noise probability (λ) as the proportions of the total instances in the image that are perturbed by noise. Specifically, the noise parameters are set as $\{(\phi, \zeta, \lambda)\} = \{(0.1, 0.0, 0.3),\ (0.2, 0.05, 0.5),\ (0.3, 0.1, 0.7)\}$.

5.2.2 Baselines.

We compare our method against several baselines, including cross-entropy loss (CE), noisy-label learning approaches for classification tasks (CoT Han et al. (2018), JoCoR Wei et al. (2020), EM Bekker & Goldberger (2016)), noise-robust losses for segmentation T-Loss Gonzalez-Jimenez et al. (2025), as well as techniques that account for partial spatial correlations, namely Superpixel (SP) Li et al. (2021) and SpatialCorrect (SC) Yao et al. (2023). For each method, we adopted the hyperparameters reported in the publicly available implementations or original papers, assuming them to be the best practices for their respective approaches. Although proposed method supports multi-label segmentation, many of the baseline methods are designed for binary segmentation. Therefore, in JSRT dataset, we conducted experiments by treating each class (Lung, Heart, Clavicle) as a separate binary segmentation task.

5.2.3 Implementation Details.

We use a U-Net Ronneberger et al. (2015) with an EfficientNet-B0 Tan & Le (2019) encoder, pre-trained on ImageNet Deng et al. (2009). The model is optimized using Adam Kingma & Ba (2014) with a learning rate of 0.001 and batch sizes of 16 (JSRT) and 32 (WHU Building). The parameters in the prior $p(\boldsymbol{\eta}) = \mathcal{N}(\boldsymbol{\eta}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and the initial values of the parameters in the variational posterior $q(\boldsymbol{\eta}) = \mathcal{N}(\boldsymbol{\eta}|\boldsymbol{m}, \boldsymbol{\Gamma})$ are set as follows. The parameters $\boldsymbol{\mu}$, the diagonal elements of $\boldsymbol{\Sigma}$, initial values of the elements of $\boldsymbol{m}$ and the diagonal elements of $\boldsymbol{\Gamma}$ are all same and denoted as μ, σ^2, m, and γ^2, respectively. These parameters were tuned using 15% of the WHU Building training set and validation sets, then fixed for all experiments, including JSRT, as follows: $\rho = 0.75$, $m = -5$, $\gamma = 1$, $\mu = -2$, and $\sigma = 1$.

5.3 Results for JSRT and WHU Building dataset

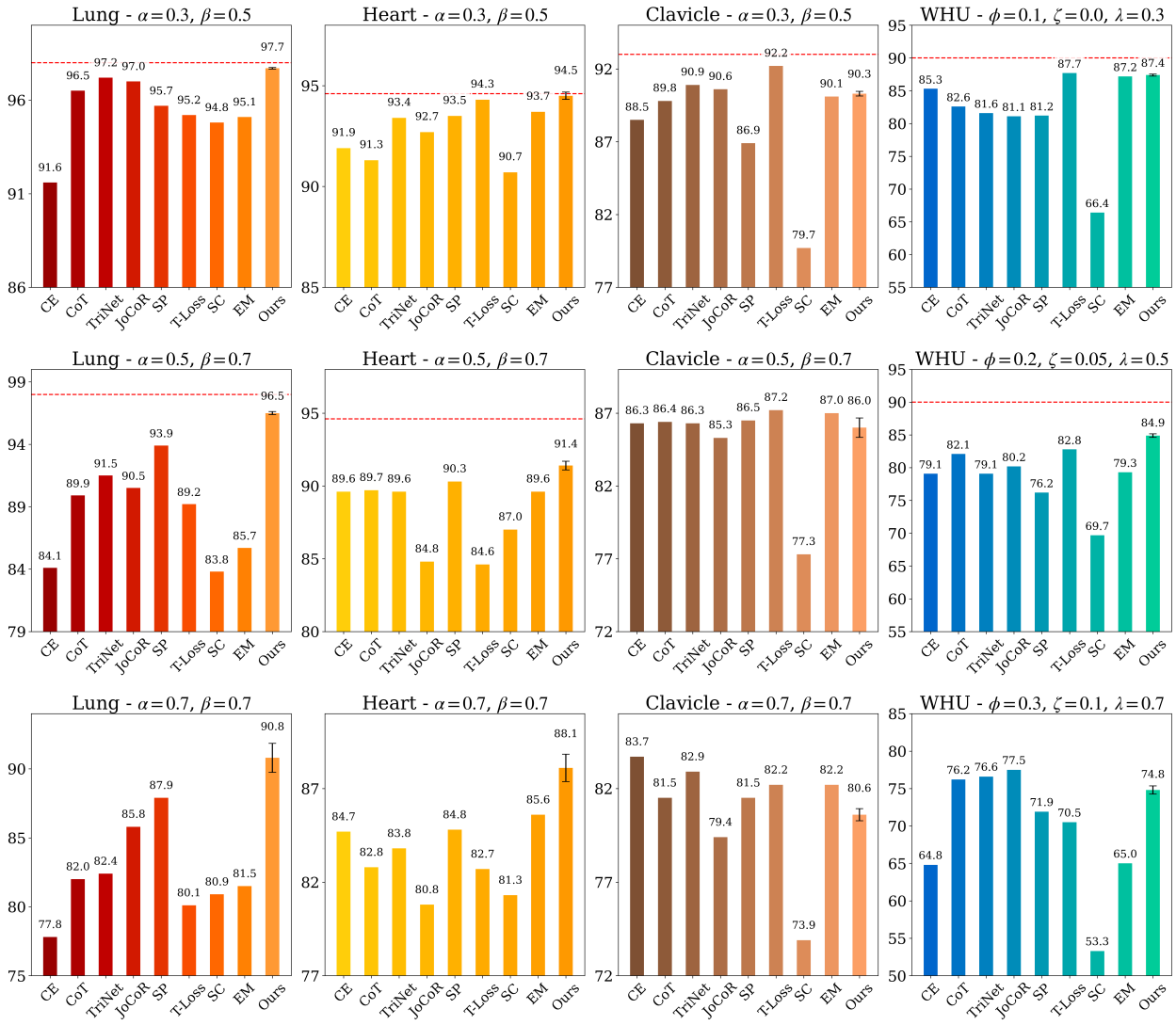


Figure 3: **Segmentation results on JSRT and WHU Building dataset.** The first left three columns show the segmentation results for three classes—Lung, Heart, and Clavicle—in the JSRT dataset, while the right column presents the binary segmentation results for WHU Building dataset in a bar chart. The rows correspond to three different noise intensity settings tested on each dataset, with Dice scores reported for JSRT and IoU scores for WHU Building dataset.

5.3.1 JSRT dataset.

As shown in the first left three columns of Figure 3, we report the Dice scores for each class (Lung, Heart, Clavicle) under all three noise conditions.

Under the moderate noise condition $(\alpha, \beta) = (0.3, 0.5)$, the CE baseline, which lacks explicit noise handling, experiences a noticeable drop of 6.4%, 2.7%, and 4.5% for each organ compared to learning with clean labels. On the other hand, most noise-robust methods exhibit only minor declines. Our proposed approach, in particular, shows a very small performance degradation in Lung (only 0.3%), indicating its strong robustness to mild noise.

As the noise level increases to $(\alpha, \beta) = (0.5, 0.7)$, CE undergoes a further 13.9% performance drop, and while other baselines also suffer more substantial declines, they remain relatively more stable than CE. Notably, our method continues to perform well on the Lung and Heart classes, demonstrating a limited reduction from the noise-free scenario. For the Clavicle class—an anatomically smaller and more challenging structure—both CE and some of the noise-robust methods experience larger drops, underscoring the sensitivity of thin or complex boundaries to label perturbations.

Under the heavier noise conditions $(\alpha, \beta) = (0.7, 0.7)$, all methods face considerable accuracy degradation. Even our approach, which remains competitive or superior for the Lung and Heart, shows more pronounced drops in Clavicle segmentation. This result highlights that extremely noisy annotations—particularly for small or intricate regions—pose significant challenges. Nevertheless, compared to the other baselines, our method tends to retain higher Dice scores and lower variance in most cases.

5.3.2 WHU Building dataset.

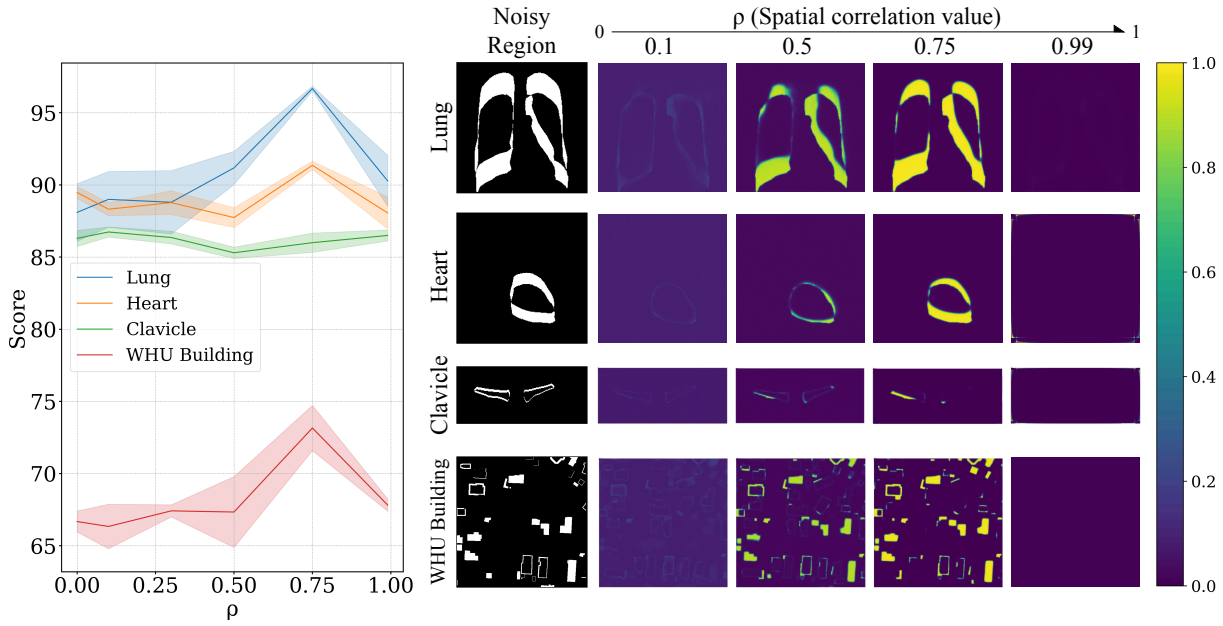
In the right column of Figure 3, we report IoU scores under three noise settings characterized by different omission, commission, and boundary probabilities (ϕ, ζ, λ). For moderate noise $(\phi = 0.1, \zeta = 0.0, \lambda = 0.3)$ and $(\phi = 0.2, \zeta = 0.05, \lambda = 0.5)$, CE's performance declines by 4.7% and 10.9% from the upper bound, respectively. Methods like CoT, TriNet, JoCoR, and SP generally fall into the low- to mid-80% range. T-Loss and EM display a smaller gap from the clean-label upper bound, reflecting their robust design. Notably, our method reaches 87.4% for $(\phi = 0.1, \zeta = 0.0, \lambda = 0.3)$, just 2.6% below the upper bound, and continues to outperform other baselines at the next noise level.

When the noise becomes stronger $(\phi = 0.3, \zeta = 0.1, \lambda = 0.7)$, all methods exhibit further deterioration, although CoT, TriNet, JoCoR, and SP still maintain relatively higher IoU scores. Our method remains competitive, indicating its adaptability even under substantial label corruption.

The results on the JSRT and WHU Building datasets confirm that our method effectively handles various types of label noise. In our experimental setting, this gives SC a disadvantage, as it is designed specifically for boundary-related biases.

5.4 Effect of Spatial Correlation Parameter ρ

Figure 4a shows how segmentation accuracy varies with the spatial correlation parameter ρ for JSRT Lung, JSRT Heart, JSRT Clavicle, and WHU Building datasets. In the Lung, Heart, and WHU Building datasets, performance peaks around $\rho = 0.75$, suggesting that moderate correlation effectively captures spatial structure and improves segmentation. In contrast, both ignoring spatial dependencies ($\rho = 0$) and assuming near-complete correlation ($\rho = 0.99$) result in noticeable accuracy drops. Performance on the Clavicle class differs, remaining relatively stable from $\rho = 0$ to 0.75, likely due to its smaller, elongated shape, which reduces the benefits of spatial correlation. Figure 4b further illustrates the estimated posterior label error probability in selected noisy regions under different ρ values. At $\rho = 0.75$, mislabeled pixels are identified more accurately, indicating that the model's label-correction mechanism is most effective under moderate correlation. In contrast, pushing ρ too high (e.g., 0.99) makes optimization unstable, leading to less reliable noise correction. Overall, these findings suggest that incorporating an appropriate level of spatial correlation significantly enhances the label correction mechanism.



(a) (b)

Figure 4: **(a) Effect of ρ.** Segmentation performance was evaluated across datasets based on ρ, which represents the degree of spatial correlation, affects segmentation performance across datasets. The results show that incorporating moderate spatial correlation enhances segmentation performance. Dice scores are reported for the Lung, Heart, and Clavicle, and IoU is used for the WHU Building dataset.
(b) Estimated label error. This figure visualizes the label error probability, computed as the sigmoid of $\boldsymbol{m}$, the posterior mean of the logit of label error $\boldsymbol{\eta}$. As shown in (a), higher accuracy in estimating $\boldsymbol{\eta}$ leads to better segmentation performance.

6 Conclusion

We have proposed a Bayesian framework for segmentation from noisy labels, introducing the logit of the label error probability as a continuous latent variable, $\boldsymbol{\eta}$. By incorporating the KMS matrix into the covariance structure of both the prior and the variational posterior distributions, our method efficiently computes the ELBO while capturing spatial correlations among adjacent pixels. This avoids the need to enumerate label combinations, improving computational efficiency. Our approach overcomes key limitations of conventional methods that assume pixel-wise independence or only partially model spatial correlation. Experiments on medical imaging and remote sensing datasets demonstrate superior robustness to noisy annotations when the spatial correlation parameter is set to a moderate value (e.g., $\rho = 0.75$).

The **ECCD** we introduced provides a new, powerful tool for probabilistic modeling with applications extending far beyond 2D segmentation. It offers a tractable framework for problems where priors on correlated discrete variables are defined over structured domains, such as spatial grids or temporal sequences. Potential avenues for future work include extending this framework to higher-dimensional spatial problems, like 3D segmentation and volumetric data analysis. Furthermore, as the underlying KMS matrix formulation is naturally suited for modeling exponentially decaying dependencies, the ECCD holds significant promise for domains beyond computer vision, particularly for modeling temporal correlations in time-series analysis and other sequential data.

Our method has some limitations. First, storing and optimizing posterior parameters for each training sample increases computational overhead though this cost is comparable to methods that require maintaining multiple models in memory. Second, spatial correlation may be less beneficial when the target structures are small or elongated, as the spatial dependencies may not contribute effectively.

Broader Impact Statement

This work introduces the ECCD, a framework designed to enhance the robustness of machine learning models against spatially correlated label noise. We believe this research has the potential for significant positive societal impact, particularly in domains where high-quality data annotation is a major bottleneck. In medical imaging, our method could lower the barrier to developing diagnostic support tools by enabling the use of larger, more diverse, and imperfectly labeled datasets, potentially accelerating research and improving healthcare accessibility. Similarly, in remote sensing, the ECCD can facilitate more accurate and large-scale analysis of satellite imagery for applications such as environmental monitoring, disaster response, and urban planning, by effectively handling noise inherent in large-scale data collection.

At the same time, we recognize potential negative repercussions that users of this technology should consider. In high-stakes applications like medical diagnosis, over-reliance on the output of any automated system, including one trained with our method, can lead to automation bias and potential misdiagnosis if not used with expert oversight. We emphasize that this technology should be used as a decision-support tool to assist qualified professionals, not to replace them. Furthermore, like many remote sensing technologies, improved segmentation capabilities could be repurposed for surveillance or other applications with negative societal consequences. We advocate for the ethical use of this technology in accordance with established legal and ethical guidelines. Finally, the performance of our method relies on the assumption of stationary, exponentially decaying correlations in the label noise. Users should be aware that in scenarios where this assumption does not hold, the model's effectiveness may be reduced.

References

Alan Joseph Bekker and Jacob Goldberger. Training Deep Neural-Netorks Based on Unreliable Labels. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2682–2686, 2016.

Antonin Berthon, Bo Han, Gang Niu, Tongliang Liu, and Masashi Sugiyama. Confidence Scores Make Instance-Dependent Label-Noise Learning Possible. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 825–836, 2021.

Peter J. Brockwell and Richard A. Davis. *Time Series: Theory and Methods*. Springer Series in Statistics. Springer-Verlag New York, 1991.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

George Fikioris. Spectral Properties of Kac–Murdock–Szegö Matrices With a Complex Parameter. *Linear Algebra and its Applications*, 553:182–210, 2018.

Giles M. Foody. Status of Land Cover Classification Accuracy Assessment. *Remote Sensing of Environment*, 80:185–201, 2002.

Benoît Frénay and Michel Verleysen. Classification in the Presence of Label Noise: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2013.

Steffen Fritz, Ian Mccallum, Christian Schill, Christoph Perger, Roland Grillmayer, Frédéric Achard, Florian Kraxner, and Michael Obersteiner. Geo-Wiki.Org: The Use of Crowdsourcing to Improve Global Land Cover. *Remote. Sens.*, 1:345–354, 2009.

Alvaro Gonzalez-Jimenez, Simone Lionetti, Philippe Gottfrois, Fabian Gröger, Alexander Navarini, and Marc Pouly. Robust T-Loss for Medical Image Segmentation. *Medical Image Analysis*, 105:103735, 2025. doi: 10.1016/j.media.2025.103735. URL `https://doi.org/10.1016/j.media.2025.103735`.

Ulf Grenander and Gabor Szegö. *Toeplitz Forms and Their Applications*. University of California Press, 1958.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust Training of Deep Neural Networks with Extremely Noisy Labels. *Advances in Neural Information Processing Systems*, 31, 2018.

Corentin Henry, Friedrich Fraundorfer, and Eleonora Vig. Aerial Road Segmentation in the Presence of Topological Label Noise. In *Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 2336–2343. IEEE, 2021.

Zhongling Huang, Corneliu Octavian Dumitru, Zongxu Pan, Bin Lei, and Mihai Datcu. Classification of Large-Scale High-Resolution SAR Images With Deep Transfer Learning. *IEEE Geoscience and Remote Sensing Letters*, 18(1):107–111, 2021.

Shunping Ji, Shiqing Wei, and Meng Lu. Fully Convolutional Networks for Multisource Building Extraction from an Open Aerial and Satellite Imagery Data Set. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1):574–586, 2018.

Zhe Jiang, Wenchong He, Marcus Kirby, Sultan Asiri, and Da Yan. Weakly Supervised Spatial Deep Learning Based on Imperfect Vector Labels With Registration Errors. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 767–775, 2021.

Laura Gálvez Jiménez and Christine Decaestecker. Impact of Imperfect Annotations on CNN Training and Performance for Instance Segmentation and Classification in Digital Pathology. *Computers in Biology and Medicine*, 177:108586, 2024.

Marek Kac, W. L. Murdock, and Gabor Szegö. On the Eigen-Values of Certain Hermitian Forms. *Journal of Rational Mechanics and Analysis*, 2:767–800, 1953.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

Thomas A. Lampert, André Stumpf, and Pierre Gâncarski. An Empirical Study Into Annotator Agreement, Ground Truth Estimation, and Algorithm Evaluation. *IEEE Transactions on Image Processing*, 25(6): 2557–2572, 2016.

Shuailin Li, Zhitong Gao, and Xuming He. Superpixel-Guided Iterative Learning from Noisy Labels for Medical Image Segmentation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI 2021)*, pp. 525–535. Springer, 2021.

Chenying Liu, Conrad M. Albrecht, Yi Wang, Qingyu Li, and Xiao Xiang Zhu. AIO2: Online correction of object labels for deep learning with incomplete annotation in remote sensing image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

Sheng Liu, Kangning Liu, Weicheng Zhu, Yiqiu Shen, and Carlos Fernandez-Granda. Adaptive Early-Learning Correction for Segmentation from Noisy Annotations. *CVPR 2022*, 2022.

Yinhe Liu, Yanfei Zhong, Ailong Ma, Ji Zhao, and Liangpei Zhang. Cross-resolution national-scale land-cover mapping based on noisy label learning: A case study of china. *International Journal of Applied Earth Observation and Geoinformation*, 118:103265, 2023.

Volodymyr Mnih and Geoffrey E. Hinton. Learning to Label Aerial Images from Noisy Data. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pp. 567–574, 2012.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2233–2241, 2017.

Charlotte Pelletier, Silvia Valero, Jordi Inglada, Nicolas Champion, Claire Marais-Sicre, and Gérard Dedieu. Effect of Training Class Label Noise on Classification Performances for Land Cover Mapping with Satellite Image Time Series. *Remote Sensing*, 9:173, 2017.

Tim Rädsch, Annika Reinke, Vivienn Weru, Minu D. Tizabi, Nicholas Schreck, A. Emre Kavur, Bünyamin Pekdemir, Tobias Roß, Annette Kopp-Schneider, and Lena Maier-Hein. Labelling Instructions Matter in Biomedical Image Analysis. *Nature Machine Intelligence*, 5(3):273–283, 2023.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI 2015)*, pp. 234–241. Springer, 2015.

Jialin Shi and Ji Wu. Distilling Effective Supervision for Robust Medical Image Segmentation with Noisy Labels. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 668–677. Springer, 2021.

Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. Development of a Digital Image Database for Chest Radiographs With and Without a Lung Nodule: Receiver Operating Characteristic Analysis of Radiologists' Detection of Pulmonary Nodules. *American Journal of Roentgenology*, 174(1): 71–74, 2000.

Hunsoo Song, Lexie Yang, and Jinha Jung. Self-filtered Learning for Semantic Segmentation of Buildings in Remote Sensing Imagery With Noisy Labels. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 16:1113–1129, 2022.

Charles Sutton and Andrew McCallum. An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373, 2012.

Mingxing Tan and Quoc Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *International Conference on Machine Learning*, pp. 6105–6114. PMLR, 2019.

Şerban Vădineanu, Daniël Maria Pelt, Oleh Dzyubachyk, and Kees Joost Batenburg. An Analysis of the Impact of Annotation Errors on the Accuracy of Deep Learning for Cell Segmentation. In *International Conference on Medical Imaging with Deep Learning*, pp. 1251–1267. PMLR, 2022.

Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating Noisy Labels by Agreement: A Joint Training Method with Co-Regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Junshi Xia, Naoto Yokoya, Bruno Adriano, and Clifford Broni-Bediako. OpenEarthMap: A Benchmark Dataset for Global High-Resolution Land Cover Mapping. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6254–6264, 2023.

Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent Label Noise: Towards Instance-Dependent Label Noise. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7597–7610, 2020.

Cheng Xue, Qiao Deng, Xiaomeng Li, Qi Dou, and Pheng-Ann Heng. Cascaded Robust Learning at Imperfect Labels for Chest X-ray Segmentation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI 2020)*, pp. 579–588. Springer, 2020.

Feng Yang, Ghada Zamzmi, Sandeep Angara, Sivaramakrishnan Rajaraman, André Aquilina, Zhiyun Xue, Stefan Jaeger, Emmanouil Papagiannakis, and Sameer K. Antani. Assessing Inter-Annotator Agreement for Medical Image Segmentation. *IEEE Access*, 11:21300–21312, 2023.

Jiachen Yao, Yikai Zhang, Songzhu Zheng, Mayank Goswami, Prateek Prasanna, and Chao Chen. Learning to Segment from Noisy Annotations: A Spatial Correction Approach. In *International Conference on Learning Representations*, 2023.

Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, and Masashi Sugiyama. Dual T: Reducing Estimation Error for Transition Matrix in Label-Noise Learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7260–7271, 2020.

Le Zhang, Ryutaro Tanno, Mou-Cheng Xu, Chen Jin, Joseph Jacob, Olga Cicarrelli, Frederik Barkhof, and Daniel Alexander. Disentangling Human Error from Ground Truth in Segmentation of Medical Images. *Advances in Neural Information Processing Systems*, 33:15750–15762, 2020a.

Mengmeng Zhang, Wei Li, Ran Tao, and Song Wang. Transfer Learning for Optical and SAR Data Correspondence Identification With Limited Training Labels. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:1545–1557, 2021.

Minqing Zhang, Jiantao Gao, Zhen Lyu, Weibing Zhao, Qin Wang, Weizhen Ding, Sheng Wang, Zhen Li, and Shuguang Cui. Characterizing Label Errors: Confident Learning for Noisy-Labeled Image Segmentation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI 2020)*, pp. 721–730. Springer, 2020b.

Tianwei Zhang, Lequan Yu, Na Hu, Su Lv, and Shi Gu. Robust Medical Image Segmentation from Non-Expert Annotations with Tri-Network. In *Medical Image Computing and Computer Assisted Intervention (MICCAI 2020)*, pp. 249–258. Springer, 2020c.

Appendix

A Derivation details

In this section, we expand on several details that were only briefly mentioned in the main text. First, we provide a comprehensive explanation of the KMS matrix (see Section A.1). Next, we outline the intermediate steps in the derivation of the ELBO in our method and the derivation of the ELBO terms that account for spatial correlation (see Section A.2). We also derive the determinant and inverse of the covariance matrices $\mathbf{\Sigma}$ and $\mathbf{\Gamma}$, which are constructed using the KMS covariance (see Section A.3). In particular, Section A.3 details how we exploit the sparse structure of these inverses to achieve efficient computations.

A.1 KMS Matrix

The *Kac–Murdock–Szegö (KMS) matrix* Fikioris (2018); Kac et al. (1953) is defined for a positive integer $n \in \mathbb{N}$ and a parameter $\rho \in (-1, 1)$ as

$$A_n(\rho) := \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \rho^2 & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \rho & \cdots & \rho^{n-3} \\ \rho^3 & \rho^2 & \rho & 1 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \rho \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & \rho & 1 \end{bmatrix} \in \mathbb{R}^{n \times n}. \tag{15}$$

The (i, j) element of the KMS matrix is given by $\rho^{|i-j|}$, capturing the idea that the correlation between elements decays exponentially with the distance between indices.

Determinant of $A_n(\rho)$

To derive the determinant, we start with

$$\det A_n(\rho) = \begin{vmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{vmatrix}. \tag{16}$$

For $i = 2, \ldots, n$, subtract ρ^{i-1} times the first row from the ith row. This operation zeroes out the first column (except the first row) and introduces factors of $(1-\rho^2)$ in the diagonal of the resulting submatrix. Proceeding recursively, one obtains

$$\det A_n(\rho) = (1-\rho^2)^{n-1}. \tag{17}$$

This compact expression is key for efficient computation.

Inverse of $A_n(\rho)$

The inverse is given by

$$A_n(\rho)^{-1} = \frac{1}{\det A_n(\rho)} \operatorname{adj}(A_n(\rho)), \tag{18}$$

where the adjugate $\operatorname{adj}(A_n(\rho))$ is the transpose of the cofactor matrix. Through a series of elementary row operations and by exploiting the Toeplitz structure, it can be shown that the cofactors also share a patterned structure. In particular, one finds that

$$\operatorname{adj}(A_n(\rho)) = (1-\rho^2)^{n-2}\, B_n, \tag{19}$$

with

$$B_n = \begin{bmatrix} 1 & -\rho & 0 & \cdots & 0 \\ -\rho & 1+\rho^2 & -\rho & \cdots & 0 \\ 0 & -\rho & 1+\rho^2 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & -\rho \\ 0 & 0 & \cdots & -\rho & 1 \end{bmatrix}. \tag{20}$$

Thus, using $\det A_n(\rho) = (1-\rho^2)^{n-1}$, we obtain

$$A_n(\rho)^{-1} = \frac{(1-\rho^2)^{n-2}}{(1-\rho^2)^{n-1}}\, B_n = \frac{1}{1-\rho^2}\, B_n. \tag{21}$$

These derivations are essential in our framework for efficiently computing the covariance matrix inverse and determinant used in the Bayesian model.

A.2 Derivation of ELBO (Eq.(3))

We assume the following generative model for our segmentation task:

$$p(\mathbf{y}^*, \mathbf{y}, \boldsymbol{\eta}|\mathbf{x}) = p(\mathbf{y}^*|\mathbf{x})\, p(\mathbf{y}|\mathbf{y}^*, \boldsymbol{\eta})\, p(\boldsymbol{\eta}) \tag{22}$$

$$= \Big\{\prod_i p(y_i^*|\mathbf{x})\Big\} \Big\{\prod_j p(y_j|y_j^*, \eta_j)\Big\}\, p(\boldsymbol{\eta}), \tag{23}$$

where the prior $p(\boldsymbol{\eta})$ is designed to capture spatial correlation (via a KMS covariance structure).

We introduce the variational distribution as

$$:= q(\boldsymbol{\eta}, \mathbf{y}^*|\mathbf{y}) = q(\boldsymbol{\eta}|\mathbf{y})\, q(\mathbf{y}^*|\boldsymbol{\eta}, \mathbf{y}) = q(\boldsymbol{\eta}) \prod_i q(y_i^*|\eta_i, y_i), \tag{24}$$

which approximates the true posterior $p(\mathbf{y}^*, \boldsymbol{\eta}|\mathbf{y}, \mathbf{x})$. Then the marginal log-likelihood of the observed labels can be lower bounded as follows:

$$\log p(\mathbf{y}|\mathbf{x}) = \log \int \sum_{\mathbf{y}^*} p(\mathbf{y}|\boldsymbol{\eta}, \mathbf{y}^*, \mathbf{x})\, p(\boldsymbol{\eta}, \mathbf{y}^*|\mathbf{x})\, d\boldsymbol{\eta} \tag{25}$$

$$= \log \mathbb{E}\left[\frac{p(\mathbf{y}|\boldsymbol{\eta}, \mathbf{y}^*, \mathbf{x})\, p(\boldsymbol{\eta}, \mathbf{y}^*|\mathbf{x})}{}\right] \tag{26}$$

$$\geq \mathbb{E}\left[\log p(\mathbf{y}|\boldsymbol{\eta}, \mathbf{y}^*, \mathbf{x})\right] - KL\Big[\,\|\, p(\boldsymbol{\eta}, \mathbf{y}^*|\mathbf{x})\Big]. \tag{27}$$

We decompose the expectation term as:

$$
\begin{aligned}
&\mathbb{E}\left[\log p(\mathbf{y}|\boldsymbol{\eta}, \mathbf{y}^*, \mathbf{x})\right] && (28)\\
&= \mathbb{E}\left[\log p(\mathbf{y}|\boldsymbol{\eta}, \mathbf{y}^*, \mathbf{x})\right] && (29)\\
&= \mathbb{E}\left[\log p(\mathbf{y}|\boldsymbol{\eta}, \mathbf{y}^*)\right] && (30)\\
&= \mathbb{E}\left[\log\left(\prod_i p(y_i|\eta_i, y_i^*)\right)\right] && (31)\\
&= \mathbb{E}\left[\sum_i \log p(y_i|\eta_i, y_i^*)\right] && (32)\\
&= \sum_i \mathbb{E}\left[\log p(y_i|\eta_i, y_i^*)\right] && (33)\\
&= \sum_i \mathbb{E}_{q(\eta_i, y_i^*|\mathbf{y})}\left[\log p(y_i|\eta_i, y_i^*)\right] \quad \text{(marginalized by } (\boldsymbol{\eta}_{\setminus i}, \mathbf{y}^*_{\setminus i})) && (34)\\
&= \sum_i \mathbb{E}_{q(y_i^*|\eta_i, \mathbf{y})q(\eta_i)}\left[\log p(y_i|\eta_i, y_i^*)\right] && (35)\\
&= \sum_i \mathbb{E}_{q(y_i^*|\eta_i, y_i)q(\eta_i)}\left[\log p(y_i|\eta_i, y_i^*)\right], \quad (y_i^* \perp\!\!\!\perp \mathbf{y}_{\setminus i}|\eta_i, y_i) && (36)
\end{aligned}
$$

where $\setminus i$ denotes pixel indices other than i on the image. We also decompose the negative KL divergence term as:

$$
\begin{aligned}
&-KL\left[\|p(\boldsymbol{\eta}, \mathbf{y}^*|\mathbf{x})\right] && (37)\\
&= \mathbb{E}\left[\log \frac{p(\boldsymbol{\eta}, \mathbf{y}^*|\mathbf{x})}{}\right] && (38)\\
&= \mathbb{E}\left[\log p(\boldsymbol{\eta}, \mathbf{y}^*|\mathbf{x}) - \log\right] && (39)\\
&= \mathbb{E}\left[\log \underbrace{p(\mathbf{y}^*|\boldsymbol{\eta}, \mathbf{x})}_{=\,p(\mathbf{y}^*|\mathbf{x})} \underbrace{p(\boldsymbol{\eta}|\mathbf{x})}_{=\,p(\boldsymbol{\eta})} - \log q(\mathbf{y}^*|\mathbf{y}, \boldsymbol{\eta}) \underbrace{q(\boldsymbol{\eta}|\mathbf{y})}_{=\,q(\boldsymbol{\eta})}\right] && (40)\\
&= \mathbb{E}\left[\log \underbrace{p(\mathbf{y}^*|\mathbf{x})}_{=\,\prod_i p(y_i^*|\mathbf{x})} + \log p(\boldsymbol{\eta}) - \log \underbrace{q(\mathbf{y}^*|\mathbf{y}, \boldsymbol{\eta})}_{=\,\prod_i \{q(y_i^*|y_i, \eta_i)\}} - \log q(\boldsymbol{\eta})\right] && (41)\\
&= \mathbb{E}\left[\sum_i \log p(y_i^*|\mathbf{x}) + \log p(\boldsymbol{\eta}) - \sum_i \log q(y_i^*|y_i, \eta_i) - \log q(\boldsymbol{\eta})\right] && (42)\\
&= -\mathbb{E}_{q(\boldsymbol{\eta})}\left[\log q(\boldsymbol{\eta})\right] + \mathbb{E}_{q(\boldsymbol{\eta})}\left[\log p(\boldsymbol{\eta})\right] - \sum_i \mathbb{E}_{q(\eta_i)q(y_i^*|\eta_i, y_i)}\left[\log q(y_i^*|y_i, \eta_i)\right] && (43)\\
&\quad + \sum_i \mathbb{E}_{q(\eta_i)q(y_i^*|\eta_i, y_i)}\left[\log p(y_i^*|\mathbf{x})\right] && (44)
\end{aligned}
$$

Thus, the overall ELBO becomes

$$
\begin{aligned}
\log p(\mathbf{y}|\mathbf{x}) \geq &\sum_i \mathbb{E}_{q(\eta_i)\, q(y_i^*|y_i, \eta_i)}\left[\log p(y_i|\eta_i, y_i^*)\right] - \mathbb{E}_{q(\boldsymbol{\eta})}\left[\log q(\boldsymbol{\eta})\right] + \mathbb{E}_{q(\boldsymbol{\eta})}\left[\log p(\boldsymbol{\eta})\right]\\
&- \sum_i \mathbb{E}_{q(\eta_i)\, q(y_i^*|y_i, \eta_i)}\left[\log q(y_i^*|y_i, \eta_i)\right] + \sum_i \mathbb{E}_{q(\eta_i)\, q(y_i^*|y_i, \eta_i)}\left[\log p(y_i^*|\mathbf{x})\right]. \qquad (45)
\end{aligned}
$$

A.3 Efficient ELBO Computation

In our framework, the spatially correlated prior $p(\boldsymbol{\eta})$ is modeled as a Gaussian distribution with a covariance matrix that incorporates spatial structure via a Kac–Murdock–Szegö (KMS) matrix:

$$\boldsymbol{\Sigma} = V\left(R_\rho \otimes R'_\rho\right)V, \tag{46}$$

where $R_\rho = A_H(\rho) \in \mathbb{R}^{H\times H}$ and $R'_\rho = A_W(\rho) \in \mathbb{R}^{W\times W}$ are the KMS matrices for the vertical and horizontal directions, respectively, and V is a diagonal matrix (e.g., $V = \sigma I$ in a simplified setting).

Because the inverse of a KMS matrix is tridiagonal, the inverse of the Kronecker product $R_\rho^{-1} \otimes {R'_\rho}^{-1}$ is very sparse (with at most 9 nonzero entries per row). Hence, if $V = \sigma I$, then

$$\boldsymbol{\Sigma}^{-1} = V^{-1}\left(R_\rho^{-1} \otimes {R'_\rho}^{-1}\right)V^{-1} = \frac{1}{\sigma^2}(R_\rho^{-1} \otimes {R'_\rho}^{-1}). \tag{47}$$

Similarly, by exploiting the determinant properties of Kronecker products,

$$\begin{aligned}\det\boldsymbol{\Sigma} &= (\det V)^2\,(\det R_\rho)^W(\det R'_\rho)^H \\ &= \Big(\prod_{i=1}^{HW}\sigma_i\Big)^2(1-\rho^2)^{W(H-1)}(1-\rho^2)^{H(W-1)}.\end{aligned} \tag{48}$$

Taking the logarithm yields

$$\log\det\boldsymbol{\Sigma} = 2\sum_{i=1}^{HW}\log\sigma_i + W(H-1)\log(1-\rho^2) + H(W-1)\log(1-\rho^2). \tag{49}$$

The efficient computation of the trace terms in the ELBO is similarly achieved by noting that $\boldsymbol{\Sigma}^{-1}$ is sparse, which reduces the computational burden when evaluating expressions such as $\operatorname{tr}\big(\boldsymbol{\Gamma}\boldsymbol{\Sigma}^{-1}\big)$ and $\operatorname{tr}\big((\boldsymbol{\mu}-\boldsymbol{m})(\boldsymbol{\mu}-\boldsymbol{m})^T\,\boldsymbol{\Sigma}^{-1}\big)$.