CLC _____ Number _____

UDC _____ Available for reference ☐Yes ☐No

SUSTech Southern University of Science and Technology

Undergraduate Thesis

Thesis Title: **Model Design for Noise-Robust Object Detection under Topological Constraints**

Student Name: **Yiran CHEN**

Student ID: **12211117**

Department: **Department of Mathematics**

Program: **Mathematics**

Thesis Advisor: **Prof. Yifei ZHU**

COMMITMENT OF HONESTY

1. I solemnly promise that the paper presented comes from my independent research work under my supervisor's supervision. All statistics and images are real and reliable.
2. Except for the annotated reference, the paper contents no other published work or achievement by person or group. All people making important contributions to the study of the paper have been indicated clearly in the paper.
3. I promise that I did not plagiarize other people's research achievement or forge related data in the process of designing topic and research content.
4. If there is violation of any intellectual property right, I will take legal responsibility myself.

Signature:

Date:

Model Design for Noise-Robust Object Detection under Topological Constraints

[ABSTRACT]: Robust object detection is an important problem in computer vision because real images are often affected by noise and other imperfect conditions. In crowded scenes, noise may weaken object responses, introduce background activations, or change the separation between nearby instances. In this thesis, we study noise-robust single-class object detection with YOLO11 and explore whether topological structure can provide useful guidance beyond ordinary local response matching.

In the proposed method, we build a shared stride-8 objectness landscape from pre-NMS boxes and confidence scores. The noisy student model is then trained to match the clean teacher model in both heatmap values and landscape topology. Specifically, we use superlevel filtration and 0-dimensional persistent homology to describe the birth and merging of connected components, and combine Dice heatmap alignment with a Betti-matching topological loss. Experiments on a pedestrian detection dataset show that the hybrid Dice+Betti objective improves robustness under medium and high Gaussian noise. Moreover, visualizations of heatmaps, persistence barcodes, and persistence diagrams support the structural interpretation of the method.

[Keywords]: Robust Object Detection, YOLO11, Knowledge Distillation, Objectness Landscape, Topological Data Analysis, Persistent Homology

[摘要]：

鲁棒目标检测是计算机视觉中的重要问题，因为真实图像往往受到噪声条件影响。在行人密集或目标相互接近的场景中，噪声可能削弱真实目标响应、引入背景激活，或改变相邻目标之间的分离关系。本文以 YOLO11 单类别目标检测为研究对象，探讨拓扑结构是否能够为噪声鲁棒性提供不同于普通局部响应匹配的监督信息。

本文将 NMS 之前的解码候选框及其置信度通过 B 样条核重建到统一的 stride-8 网格上，形成目标性景观。在教师–学生知识蒸馏框架下，噪声图像上的学生模型不仅学习干净教师模型的热图数值，也学习该景观的拓扑结构。具体而言，本文利用超水平集滤波和 0 维持久同调描述连通分支的出生与合并过程，并将 Dice 热图对齐项与基于 Betti 匹配的拓扑损失结合。实验结果表明，Dice+Betti 混合目标在中高强度高斯噪声下能够提升一定的检测鲁棒性，热图、持久条形码和持久图的可视化结果也支持本文对目标性结构的解释。

[关键词]：鲁棒目标检测； YOLO11；知识蒸馏；目标性景观；拓扑数据分析；持久同调

Contents

1. Introduction

1.1 YOLO-Based Object Detection

Object detection is a core task in computer vision. One-stage detectors such as YOLO make this task efficient by treating detection as dense prediction [7]. However, dense prediction is also sensitive to noise. Noise can change feature responses and create false activations in cluttered regions.

YOLO is usually described through three modules. The backbone extracts visual features. The neck fuses features from different scales. The detection head converts the fused features into predictions. After multi-scale feature maps are produced, the head applies a final convolutional layer. For each spatial location, it predicts bounding-box parameters, class probabilities, and an objectness score.

1.2 Objectness Maps as Response Landscapes

For a given image, the objectness scores' distribution forms an objectness map. The map indicates how likely each position is to contain an object. It can also be viewed as a landscape with peaks and valleys. From this perspective, object localization becomes a structural problem.

High objectness responses form peaks near object instances. Low responses form valleys between different regions. This landscape view helps us study how noise changes the structure of objectness predictions.

1.3 Noise-Induced Distortions in Objectness Landscapes

When a pretrained detector is applied to noisy inputs, its objectness landscape may be distorted. Compared with the clean-input landscape, sharp peaks can become flatter, shallow valleys can rise, and extra fluctuations can appear. These changes affect both the strength and the separation of activation regions.

For this reason, peaks that are separate in clean images may become connected under noise. Different objects may then be merged. The opposite problem can also occur. A single clear peak in the clean landscape may split into several local maxima in the noisy landscape,

which can lead to duplicate or unstable detections.

1.4 A Topological View of Response Connectivity

These distortions indicates that the objectness landscape is vulnerable to noise. They also suggest that robustness should not be studied only at the pixel level, but also focus on the connectivity between object responses.

To capture such connectivity variation in a principled yet model-agnostic way, we adopt a topological data analysis (TDA) viewpoint [3], where the objectness map is conceptually "scanned" from high values to low values. As the threshold sweeps downward, peaks emerge as isolated regions and valleys determine when these regions merge.

By tracking the appearance and merging times of peak–valley structures, TDA gives a compact description of the landscape. It then becomes possible to quantify the changes in the objectness map before and after noise is added.

1.5 Preserving Landscape Consistency Under Noise

Based on this perspective, our goal is to preserve the essential peak–valley structure of the objectness landscape when the input is noisy. We do not rely only on pixelwise consistency. Instead, we use structural information from TDA to guide the model toward predictions that remain close to the clean reference.

2. Related Work

2.1 Object Detection (YOLO)

The YOLO family formulates object detection as a one-stage dense prediction problem [7]. In the Ultralytics implementation [9], YOLO11, also called YOLOv11, uses a backbone–neck–Detect head structure. It produces feature maps at three strides, 8, 16, and 32, which can be viewed as multi-resolution sampling of the same image domain.

At each grid cell's location, the detection head outputs class scores and distance-based box parameters. These local predictions are then decoded with grid-center anchor points into image-coordinate candidate boxes.

As for outputs, YOLO produces a finite collection of detections after decoding detection head's outputs and applying post-processing.

Each retained detection consists of an image-coordinate bounding box, a predicted class label, and a confidence score. The formal input and output spaces used in this work will be introduced later in the problem setup.

This pipeline depends on local response strength, decoded geometry, and candidate ranking. Input noise can therefore disturb the final detections, which motivates the robustness discussion in the next subsection.

2.2 Noise Robustness

Robust object detection under degraded inputs has been studied from several directions. At the data level, many methods use corruption-aware augmentation, stylization, or denoising preprocessing. Benchmark studies also show that standard detectors can degrade under image corruptions and adverse weather [5, 6].

At the objective level, some methods use robust losses or consistency regularization to keep predictions stable under perturbed inputs.

Another viewpoint adopts teacher model–student model training to preserve detector behavior under degradation.

Most existing approaches focus on consistency at the pixel, feature, or output level. They usually do not directly constrain the structural organization of detections, such as peak–valley geometry or connectivity in an objectness landscape.

2.3 Knowledge Distillation for Detection

Knowledge distillation has become a standard tool for transferring detection performance from a stronger teacher model to a lighter student model.

Existing detector-specific methods distill fine-grained feature responses [10, 11]. They also distill instance-level or response-level information [2], as well as localization logits for dense prediction heads [12]. These works show that logits, feature maps, and proposal-related responses can all provide useful supervision.

In most cases, however, the transferred knowledge is defined on activations or local responses rather than global structural behavior. In particular, topological relations among decoded detection responses are rarely treated as the distillation target.

2.4 Topology in Deep Learning

Topological data analysis, especially persistent homology [3], has increasingly been used as a structural prior in deep learning.

Topological losses can help preserve shape and connectivity in segmentation [1]. Induced matchings of persistence barcodes also provide a principled way to compute structural changes between images [8].

More recently, topological information has been explored as an extra signal in knowledge distillation [4]. Still, most studies focus on segmentation masks or general feature representations. Decoded outputs of object detectors are studied less often.

Hence, our setting applies topological regularization to a reconstructed objectness landscape derived from YOLO11 predictions. We use it to enforce structural consistency between the teacher model and student model under noise.

3. Preliminaries

3.1 Intersection over Union (IoU)

Intersection over Union (IoU) measures the overlap between a predicted bounding box and the corresponding ground-truth box.

Given a predicted bounding box B_p and a ground-truth bounding box B_{gt}, IoU is defined as the ratio between their intersection area and their union area:

$$\mathrm{IoU}(B_p, B_{gt}) = \frac{|B_p \cap B_{gt}|}{|B_p \cup B_{gt}|}.$$

The IoU value ranges from 0 to 1. A higher value indicates better alignment between the predicted box and the ground-truth box. In practice, a prediction is counted as a true positive if its IoU with the ground truth exceeds a chosen threshold.

4. Method

4.1 Problem Setup

We use **YOLO11 (YOLOv11)** [9] for single-class object detection. The training framework follows teacher model–student model distillation. Let the input image domain be

$$\mathcal{X} := \{1, \ldots, \text{Color}\}^{\text{Height} \times \text{Width}},$$

and let the output detection space be

$$\mathcal{Y} := \left\{ S \subset \mathbb{R}^4 \times \mathcal{C} \times [0,1] \;\middle|\; |S| < \infty \right\}.$$

Here $\mathcal{C}$ is the finite set of object classes. Each detection is written as a triple (b_i, c_i, s_i). Here $b_i \in \mathbb{R}^4$ is the image-coordinate bounding box, $c_i \in \mathcal{C}$ is the predicted class label, and $s_i \in [0,1]$ is the confidence score. An object detection model is then defined as a mapping

$$f : \mathcal{X} \to \mathcal{Y}, \qquad x \mapsto f(x) := \{(b_i, c_i, s_i)\}_{i=1}^{N(x)},$$

where $N(x)$ is the number of detections retained for image x.

Let the detector operate on the multi-stride feature set

$$\mathcal{S} := \{8, 16, 32\}.$$

For each stride $s \in \mathcal{S}$, let $\mathcal{I}_s$ be the set of decoded prediction indices at that stride's feature level. We define

$$\mathcal{I} := \bigsqcup_{s \in \mathcal{S}} \mathcal{I}_s$$

as the collection of all decoded prediction indices across strides. For each $i \in \mathcal{I}$, the detector first produces a decoded candidate triple (b_i, c_i, s_i). The bounding box is written as

$$b_i := (x_i^{(1)}, y_i^{(1)}, x_i^{(2)}, y_i^{(2)}) \in \mathbb{R}^4$$

in image coordinates. Given the class-logit vector $z_i \in \mathbb{R}^C$, where $C := |\mathcal{C}|$, the confidence

score and predicted class are defined as

$$s_i := \max_{1\le r\le C} \sigma(z_{i,r}) \in [0,1], \qquad c_i := \arg\max_{1\le r\le C} \sigma(z_{i,r}).$$

In the single-class setting, s_i becomes the only class probability. It can be treated as an object-presence confidence, i.e objectness. The final output $f(x)$ is obtained by applying objectness filtering and IoU-based non-maximum suppression (NMS) to the decoded candidates, then NMS results can be written as

$$\{(b_i, c_i, s_i) \mid i \in \mathcal{I},\ s_i \ge \gamma\},$$

where γ is the objectness threshold and τ is the IoU threshold used by NMS. Let $\mathcal{R}(x) \subseteq \mathcal{I}$ be the retained index set after this post-processing step. Then $N(x) := |\mathcal{R}(x)|$.

4.2 Unified Heatmap Reconstruction on a Shared Stride-8 Grid

Let f_{tea} be the teacher model well-trained on clean data. Let f_{stu} be the student model only be pretrained on noisy inputs.

Given a clean image $x^{\text{clean}} \in \mathcal{X}$ and its noise corrupted version $x^{\text{noise}} \in \mathcal{X}$, the goal is to train the student model so that its reconstructed confidence landscape studies towards that of the teacher model, as well as satisfying the original detection objective.

The teacher model and student model produce prediction tensors over the same strides. Our supervision, however, is not applied separately at each scale, we adopt a single class with $|\mathcal{C}| = 1$, the class label c_i is fixed.

For heatmap reconstruction, the decoded candidates can be reduced to the pre-NMS weighted box hypotheses $\{(b_i, s_i)\}_{i\in\mathcal{I}}$. These hypotheses are combined into one supervisory field on a shared stride-8 grid cells. All decoded predictions are projected onto the shared stride-8 grid cells

$$\Omega_8 = \{1,\dots,H_8\} \times \{1,\dots,W_8\}, \qquad H_8 = \left\lceil \frac{\text{Height}}{8} \right\rceil,\ W_8 = \left\lceil \frac{\text{Width}}{8} \right\rceil.$$

The grid cells matches the spatial resolution of the stride-8 feature map. It also lets predic-

tions from strides 8, 16, and 32 contribute to the same field.

Let $K((u, v); b_i)$ be a nonnegative box-induced kernel centered at prediction b_i and evaluated at grid location $(u, v) \in \Omega_8$.

The reconstructed heatmap is defined as a nonnegative function $H : \Omega_8 \to \mathbb{R}_{\geq 0}$:

$$H(u, v) = \sum_{i \in \mathcal{I}} w_i \, K((u, v); b_i), \qquad (u, v) \in \Omega_8,$$

where $w_i = s_i$ is the confidence weight.

Thus, decoded predictions from all scales are accumulated directly over the shared stride-8 lattice.

We adopt the cubic B-spline kernel as the box-induced kernel:

$$K_{\text{bspline}}((u, v); b_i) = \lambda \, \beta_3\left(\frac{v - m_i^x}{\rho_i^x}\right) \beta_3\left(\frac{u - m_i^y}{\rho_i^y}\right),$$

Here β_3 is the cardinal cubic B-spline basis. The point (m_i^x, m_i^y) is the center of b_i. The values (ρ_i^x, ρ_i^y) are scale parameters proportional to the box size. The factor λ normalizes the peak value to 1.

Several predictions may overlap on the same grid cells. For this reason, the raw sum H can be compressed by the monotone map

$$\widehat{H} : \Omega_8 \to [0, 1), \qquad \widehat{H}(u, v) = 1 - \exp(-H(u, v)).$$

This construction defines the heatmap reconstruction operator

$$\Phi : \{(b_i, s_i)\}_{i \in \mathcal{I}} \mapsto \widehat{H},$$

which maps pre-NMS weighted box hypotheses to the compressed stride-8 heatmap on Ω_8.

This bounded sacalar field will be used for quantifying the structural changes in distillation.

4.3 Heatmap as an Objectness Landscape Proxy

The reconstructed field $\widehat{H}$ can be interpreted as an objectness-landscape proxy. High values of $\widehat{H}$ indicate regions where decoded predictions confidence accumulates. Low values correspond to background-like areas and gaps between object hypotheses.

In this manner, peaks represent coherent object responses. Valleys represent low-confidence regions that separate nearby detections.

This proxy is useful under input corruption. When a fixed detector is applied to noise corrupted images, the landscape may show flattened peaks, lifted valleys, or false high-response regions.

Such changes are hard to describe using only final bounding boxes, but become obvious after the prediction set be shown as a scalar field.

Therefore, we adopt $\widehat{H}$ as the geometric viewpoint on which structure-aware supervision is imposed.

4.4 Topological Interpretation via Superlevel Filtration

To use persistent homology under cubical complex manner, we quantize $\widehat{H}$ into $\text{Color} := 256$ grayscale levels. Then the grid Ω_8 can be treated as an implicit two-dimensional cubical complex. To describe the topology of high-response regions, we build a superlevel filtration on this quantized heatmap.

For a given threshold $\tau \in \{1, \ldots, \text{Color}\}$, we define

$$\Omega_\tau := \{(u, v) \in \Omega_8 \mid \text{quant}_{256}(\widehat{H}(u, v)) \geq \tau\}.$$

As τ decreases, these sets form a nested inclusion sequence

$$\Omega_{\tau_1} \subseteq \Omega_{\tau_2} \quad \text{for} \quad \tau_1 > \tau_2,$$

which gives the superlevel filtration over the reconstructed objectness landscape.

The resulting topological evolution can be summarized by the persistence barcode

$$\mathcal{B}_\tau = \{\, [b_\tau, d_\tau)\},$$

where each interval $[\tau_b, \tau_d)$ records the birth and death of a topological feature across threshold levels [3].

Equivalently, we can rewrite the persistence barcode as corresponding persistence diagram

$$\mathrm{Dgm}(\widehat{H}) = \{\, (b_\tau, d_\tau) \mid [b_\tau, d_\tau) \in \mathcal{B}_\tau \,\} \subset \mathbb{R}^2.$$

Here, we only focus on 0-dimensional persistent homology. Each interval corresponds to a connected component that appears at a local maximum, whiel it later merges with another component as the threshold decreases.

This persistence diagram provides a detailed description of peak strength and connectivity in the reconstructed heatmap.

4.5 Distillation Goal and Loss Function

Recent work in image segmentation shows that topology can provide an additional structural signal for knowledge distillation beyond feature alignment [4]. Denote above reconstruction operator as Φ.

Let P^{tea} and P^{stu} be the teacher model and student model pre-NMS weighted box hypotheses.

Define

$$\widehat{H}^{\text{tea}} := \Phi(P^{\text{tea}}), \qquad \widehat{H}^{\text{stu}} := \Phi(P^{\text{stu}}).$$

By construction, both $\widehat{H}^{\text{tea}}$ and $\widehat{H}^{\text{stu}}$ are bounded scalar fields on the same domain Ω_8.

We can compare their topological structures using the H_0 persistence diagrams from the superlevel filtration.

Following Stucki et al. [8], let

$$\tau(\widehat{H}^{\text{stu}}, \widehat{H}^{\text{tea}}) : \mathrm{Dgm}_0(\widehat{H}^{\text{stu}}) \to \mathrm{Dgm}_0(\widehat{H}^{\text{tea}})$$

denote the induced Betti matching between the student model and teacher model diagrams, while unmatched points are mapped to the diagonal.

Define the topology-aware loss as

$$\mathcal{L}_{\text{topo}}(\widehat{H}^{\text{stu}}, \widehat{H}^{\text{tea}}) := \sum_{q \in \text{Dgm}_0(\widehat{H}^{\text{stu}})} 2 \left\| q - \tau(\widehat{H}^{\text{stu}}, \widehat{H}^{\text{tea}})(q) \right\|_2^2.$$

This loss is well defined because both heatmaps are bounded grayscale fields.

It gives a differentiable surrogate for preserving peak-merging behavior.

To stabilize optimization, we combine the topology-aware loss with a volumetric heatmap alignment term. The Dice term follows the implementation used in the Betti-matching loss module: `DiceLoss(sigmoid=True)(input, target)`. In this call, `input` is first passed through a sigmoid before the Dice overlap is computed; in our heatmap setting this is equivalent to applying Dice loss to the reconstructed student heatmap $\widehat{H}^{\text{stu}}$ and teacher heatmap $\widehat{H}^{\text{tea}}$. With the default MONAI DiceLoss setting, the Dice term used in our heatmap distillation can be written as

$$\mathcal{L}_{\text{dice}}(\widehat{H}^{\text{stu}}, \widehat{H}^{\text{tea}}) := 1 - \frac{2 \sum_{(u,v) \in \Omega_8} \widehat{H}^{\text{stu}}(u,v) \widehat{H}^{\text{tea}}(u,v)}{\sum_{(u,v) \in \Omega_8} \widehat{H}^{\text{stu}}(u,v) + \sum_{(u,v) \in \Omega_8} \widehat{H}^{\text{tea}}(u,v)}.$$

The implementation adds small smoothing constants for numerical stability and then averages this quantity over the batch and channel dimensions. The distillation loss is

$$\mathcal{L}_{\text{kd}}(\widehat{H}^{\text{stu}}, \widehat{H}^{\text{tea}}) := \mathcal{L}_{\text{dice}}(\widehat{H}^{\text{stu}}, \widehat{H}^{\text{tea}}) + \alpha \, \mathcal{L}_{\text{topo}}(\widehat{H}^{\text{stu}}, \widehat{H}^{\text{tea}}),$$

where α controls the contribution of the topological term in the hybrid setting. The full training loss is defined as

$$\mathcal{L}_{\text{train}} := \mathcal{L}_{\text{det}} + \lambda_{\text{kd}} \, \mathcal{L}_{\text{kd}}(\widehat{H}^{\text{stu}}, \widehat{H}^{\text{tea}}),$$

where $\mathcal{L}_{\text{det}}$ is the standard detection loss in default yolo11 setting. Let λ_{kd} be the weight of the distillation term.

This loss function will ensure the student model to preserve the teacher model's heatmap landscape's geometric structure, while remaining compatible with the original detection task.

4.6 Training Pipeline

For each training batch, the student model evaluates the noisy image x^{noise} and outputs the prediction set P^{stu}.

The teacher model is evaluated without gradient updates. It evaluates the corresponding clean image x^{clean}, and outputs the prediction set P^{tea}.

Both prediction sets are passed through the same reconstruction operator Φ, which derives the teacher model and student model heatmaps $\widehat{H}^{\text{tea}}$ and $\widehat{H}^{\text{stu}}$ on the shared stride-8 grid.

The loss combines two sources of supervision. First, the default yolo11 detection loss $\mathcal{L}_{\text{det}}$ is computed on the student model outputs. Second, the KD heatmap loss $\mathcal{L}_{\text{kd}}(\widehat{H}^{\text{stu}}, \widehat{H}^{\text{tea}})$ computes the difference between the reconstructed student's heatmap and the teacher's heatmap. The sum of these terms only updates the student model.

Following shows the training pipeline's pseudo-code:

Algorithm 1: Training Pipeline with B-spline Distillation

Input: student model f_{stu} (trainable), teacher model f_{tea} (frozen), noisy input x^{noise}, clean reference x^{clean} (or matched clean image)

Initialize: $\Phi \leftarrow$ shared stride-8 reconstruction with B-spline kernel, $\lambda_{\text{kd}} \geq 0$

for *each mini-batch* **do**

- $P^{\text{stu}} \leftarrow f_{\text{stu}}(x^{\text{noise}})$;
- $\mathcal{L}_{\text{det}} \leftarrow \mathcal{L}_{\text{det}}(P^{\text{stu}})$;
- `// teacher model forward only, no gradient`
- $P^{\text{tea}} \leftarrow f_{\text{tea}}(x^{\text{clean}})$ (no grad);
- $\widehat{H}^{\text{stu}} \leftarrow \Phi(P^{\text{stu}}), \quad \widehat{H}^{\text{tea}} \leftarrow \Phi(P^{\text{tea}})$;
- `// if enabled: map compression` $\widehat{H} = 1 - \exp(-H)$ `inside` Φ
- $\mathcal{L}_{\text{kd}}(\widehat{H}^{\text{stu}}, \widehat{H}^{\text{tea}}) \leftarrow$ Dice / Betti / Dice+Betti;
- $\mathcal{L}_{\text{train}} \leftarrow \mathcal{L}_{\text{det}} + \lambda_{\text{kd}}\mathcal{L}_{\text{kd}}(\widehat{H}^{\text{stu}}, \widehat{H}^{\text{tea}})$;
- update only θ_{stu} using $\nabla_{\theta_{\text{stu}}}\mathcal{L}_{\text{train}}$;

5. Experiments

5.1 Dataset

We evaluate our method on the Social Distancing Monitoring dataset (v3) from Roboflow Universe. The dataset is released under the CC BY 4.0 license.

This dataset focuses on pedestrians. It contains many dense scenes, where objects are close to each other and overlap partially. This makes it suitable for studying object separation under noise.

To match our formulation, we use single-class detection setting. All objects are treated as one category. This simplification is consistent with our topology-aware distillation problem setting, which studies the structure of the objectness map rather than class-specific instances.

The dataset was chosen by two considerations. First, dense object layouts create challenges. In these cases, noise may change the connectivity of objectness responses, which directly reflects topological distortion. Second, the dataset size is 400 pictures in total, which is moderate. It can be trained efficiently on a single GPU, so the experiments remain practical.

5.2 Implementation Details

All experiments are based on the Ultralytics implementation of YOLO11 (YOLOv11) with PyTorch. All object categories are merged into one class. The full teacher model–student model reconstruction and distillation pipeline follows the method defined in Section 3.

To simulate corrupted conditions, we apply Gaussian noise to the input images. The noise level is controlled by the standard deviation σ. We test multiple noise intensities to evaluate robustness.

Teacher model: set to be trained follow yolo11m's default setting on above clean dataset for 100 epochs from scratch.

For heatmap distillation, we evaluate three choices for $\mathcal{L}_{\text{kd}}$: Pure training, Dice, and Hybird Dice+Betti.

Pretrain eval i.e student model without distillation: set to be trained follow yolo11n's default setting on above noisy dataset for 20 epochs from scratch.

No-KD fine-tune: set to be trained follow yolo11n's default setting on above noisy dataset for 20 epochs from Pretrain eval checkpoint.

Dice KD i.e Dice Distillation: set to be trained follow yolo11n's default setting on above noisy dataset for 20 epochs from above student model checkpoint. Distillation weight is set to be $\lambda_{\text{kd}} = 3.0$.

Dice+Betti i.e Hybrid Dice+Betti Distillation: set to be trained follow yolo11n's default setting on above noisy dataset for 20 epochs from above student model checkpoint. Distillation weight is set to be $\lambda_{\text{kd}} == 2.0$. Topology weight is set to be $\alpha = 0.001$.

Training is conducted on a single NVIDIA RTX 5090 GPU with an AMD 9800X3D CPU, using the Adam optimizer with an initial learning rate of 10^{-3}. Input images are resized to 640×640. Standard data augmentation is off. Batch size is set to be 32.

5.3 Evaluation Metrics

To evaluate detection performance, we use standard object detection metrics including Precision, Recall, and mean Average Precision (mAP).

Precision and Recall Let TP, FP, and FN denote the numbers of true positives, false positives, and false negatives. Precision and Recall are defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

A predicted bounding box is a true positive if its Intersection over Union (IoU) with a ground-truth box is above a chosen threshold.

Average Precision (AP) Average Precision summarizes the Precision–Recall curve:

$$\text{AP} = \int_0^1 \text{Precision}(r)dr$$

where r denotes recall.

mAP@0.5 The mean Average Precision at IoU threshold 0.5 is defined as:

$$\text{mAP@50} = \frac{1}{C}\sum_{c=1}^{C} \text{AP}_c^{(\text{IoU}\geq 0.5)}$$

where C is the number of classes. In our single-class setting, this becomes mAP@50 = AP.

mAP@0.5:0.95 To evaluate localization accuracy and separation ability more comprehensively, we adopt the mean Average Precision over multiple IoU thresholds. Specifically, mAP@0.5:0.95 averages AP over thresholds from 0.5 to 0.95 with a step size of 0.05 is defined as:

$$\text{mAP@0.5} : 0.95 = \frac{1}{T}\sum_{t\in\mathcal{T}} \text{mAP@}t,$$

where $\mathcal{T} = \{0.50, 0.55, 0.60, \ldots, 0.95\}$ and T is the number of thresholds. Each mAP_t is computed in the same manner as mAP@50, but with IoU threshold t.

5.4 Quantitative Results

We report the clean teacher at $\sigma = 0$ as an upper bound, together with the student settings defined above: *pretrain20_eval*, No-KD fine-tune, Dice KD, and Dice+Betti KD.

All evaluation results are reported under correspoinding dataset with increasing Gaussian noise.

Table 1. Precision under different Gaussian noise levels.

Method	$\sigma = 0$	$\sigma = 20$	$\sigma = 40$	$\sigma = 60$	$\sigma = 80$
Teacher model	0.900	–	–	–	–
Pretrain eval	0.819	0.801	0.585	0.219	0.076
No-KD fine-tune	0.880	0.865	0.855	0.836	0.837
Dice KD	0.874	0.863	0.829	0.835	0.848
Dice+Betti KD	0.872	0.853	0.848	0.858	0.849

Table 2. Recall under different Gaussian noise levels.

Method	$\sigma = 0$	$\sigma = 20$	$\sigma = 40$	$\sigma = 60$	$\sigma = 80$
Teacher model	0.892	–	–	–	–
Pretrain eval	0.643	0.611	0.349	0.131	0.059
No-KD fine-tune	0.790	0.780	0.753	0.742	0.745
Dice KD	0.793	0.784	0.783	0.736	0.729
Dice+Betti KD	0.796	0.811	0.778	0.767	0.759

Table 3. mAP@50 under different Gaussian noise levels.

Method	$\sigma = 0$	$\sigma = 20$	$\sigma = 40$	$\sigma = 60$	$\sigma = 80$
Teacher model	0.952	–	–	–	–
Pretrain eval	0.749	0.728	0.409	0.106	0.036
No-KD fine-tune	0.878	0.870	0.836	0.828	0.824
Dice KD	0.878	0.866	0.848	0.822	0.813
Dice+Betti KD	0.885	0.882	0.859	0.852	0.842

Table 4. mAP@50:95 under different Gaussian noise levels.

Method	$\sigma = 0$	$\sigma = 20$	$\sigma = 40$	$\sigma = 60$	$\sigma = 80$
Teacher model	0.772	–	–	–	–
Pretrain eval	0.400	0.389	0.156	0.026	0.008
No-KD fine-tune	0.560	0.543	0.513	0.500	0.488
Dice KD	0.561	0.542	0.516	0.495	0.479
Dice+Betti KD	0.565	0.553	0.530	0.515	0.500

The four tables show a consistent overall pattern. Our method improves noisy-input detection over the original student baseline.

Among the tested variants, the Dice+Betti hybrid gives the strongest overall mAP across noise levels.

Compared with direct noisy evaluation, all adapted student models are much more robust. The hybrid variant is also closest to the teacher model in both mAP@50 and mAP@50:95.

This result suggests that the topology-aware term is useful when it is combined with ordinary volumetric alignment.

Three numerical trends are worth noting. First, the gains are small at low noise, where the baseline is already strong. At $\sigma = 0$, Dice+Betti improves mAP@50 only from 0.878 to 0.885, and mAP@50:95 from 0.560 to 0.565.

Second, the benefit becomes clearer at medium and high noise. At $\sigma = 60$, the hybrid improves mAP@50 from 0.828 to 0.852 and mAP@50:95 from 0.500 to 0.515. At $\sigma = 80$, it improves mAP@50 from 0.824 to 0.842 and mAP@50:95 from 0.488 to 0.500.

Third, the hybrid is more stable than Dice alone. Dice KD is slightly worse than the student baseline in most mAP entries, while Dice+Betti recovers this drop and exceeds the baseline. This is consistent with the role of the topology term as a structural regularizer rather than a replacement for volumetric alignment.

Some trade-offs remain. Under low noise, Dice+Betti does not always improve precision or recall over the baseline, even when mAP improves. These cases are included because they clarify where topology-aware supervision is most useful: it helps most when corruption is strong enough to disturb object-level structure severely.

5.5 Visualization

We first inspect the optimization process. Figure 1 shows representative training curves for the Dice+Betti distillation setting and the No-KD fine-tuning baseline. The Dice+Betti run includes the additional KD loss curves, while the No-KD run only optimizes the standard detection objective.

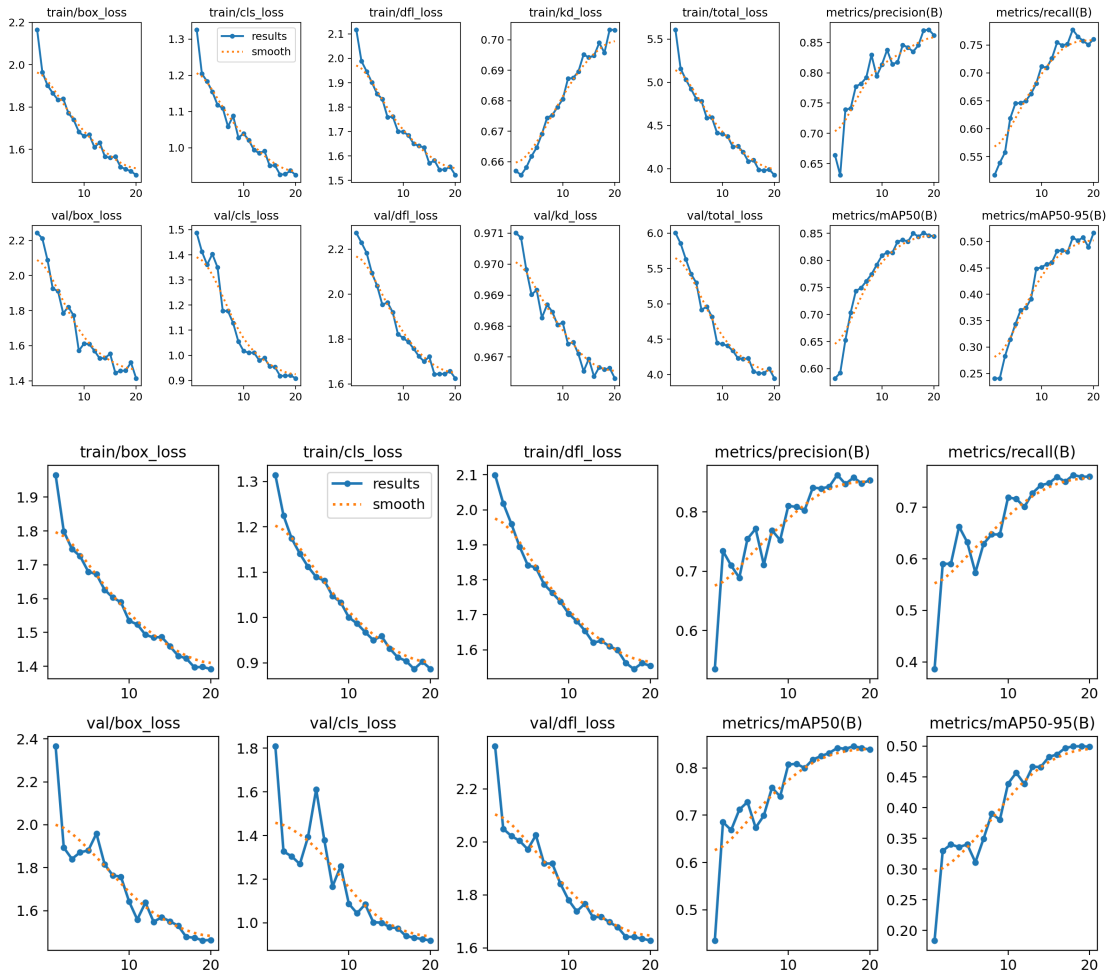

Figure 1. Training curves for Dice+Betti distillation and No-KD fine-tuning. The upper plot shows the Dice+Betti training process with KD loss terms, while the lower plot shows the No-KD fine-tuning baseline.

The training curves provide an optimization-level view of the same effect. In both settings, the ordinary detection losses decrease and validation metrics improve during fine-tuning. Compared with No-KD fine-tuning, Dice+Betti distillation starts improving earlier and shows smoother metric curves.

Its validation KD loss also decreases while the detection metrics continue to rise, indicating that the model is moving its reconstructed response field closer to the teacher model.

This trend helps explain the quantitative improvement under medium and high noise. No-KD fine-tuning improves robustness through ordinary noisy training, but it lacks explicit structural guidance. Dice+Betti distillation adds this guidance, so the later increase in mAP is consistent with the heatmap, barcode, and persistence-diagram observations.

To better understand how noise affects the clean-pretrained student model, Figure 2 shows the *pretrain20_eval* model on the same image under four noise levels.

The first row shows the final detections. The second row overlays the reconstructed B-spline heatmap on the corresponding noisy image.

Figure 2. Detection outputs and B-spline heatmap overlays of the clean-pretrained student model under different Gaussian noise levels. The top row shows predicted bounding boxes, and the bottom row shows the corresponding B-spline heatmap overlay.

Figure 3 further compares reconstructed B-spline heatmaps of different methods at $\sigma = 40$. The teacher model heatmap is computed on the corresponding clean image. This comparison examines whether distillation transfers more than detection accuracy. It also checks whether the teacher model's spatial organization of objectness responses is transferred in distillation process.

Under noise, the undistilled field or the purely local response field may contain weakened peaks, fragmented activations, or elevated background regions.

In contrast, the Dice+Betti student model produces a heatmap whose dominant high-response regions are closer to the teacher model field. The difference is especially clear around main object centers and low-response valleys between nearby instances.

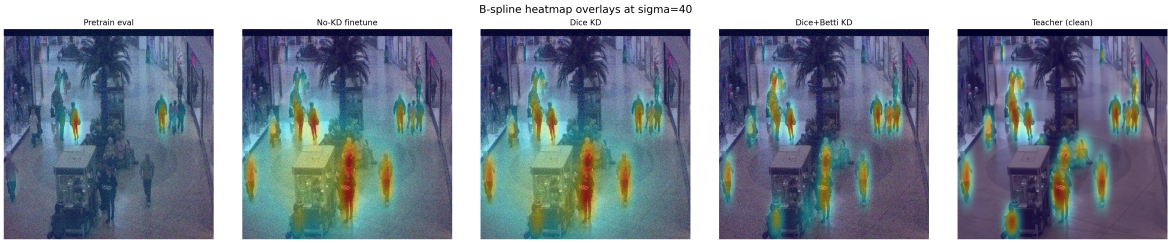

Figure 3. B-spline heatmap overlays at $\sigma = 40$ for different student model variants, with the teacher model shown on the clean image for reference. The Dice+Betti student model more closely follows the teacher model's peak–valley layout than student models trained without the topology-aware term.

To further illustrate the structural effect of the distillation process, Figures 4 and 5 show the corresponding persistence barcodes and persistence diagrams. They are computed from the same $\sigma = 40$ reconstructed fields as an example.

In the barcode view, the teacher model contains a small number of relatively persistent connected components. These components correspond to stable objectness peaks. Noisy student models tend to introduce extra short-lived bars or alter the lifetimes of important components.

After Dice+Betti distillation, the student model barcode becomes closer to the teacher model pattern. Spurious short-lived components are reduced, and the more persistent bars align better with the teacher model's dominant topological features.

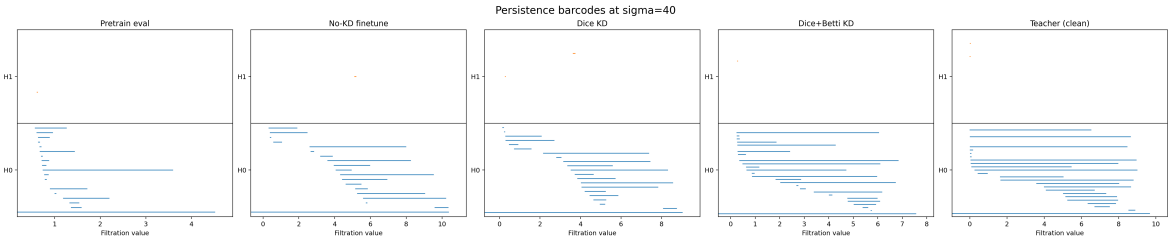

Figure 4. Persistence barcodes at $\sigma = 40$ for the compared methods, showing the lifetime of topological features across filtration values. Compared with the baseline student models, the Dice+Betti barcode better matches the teacher model in the number and persistence of dominant connected components.

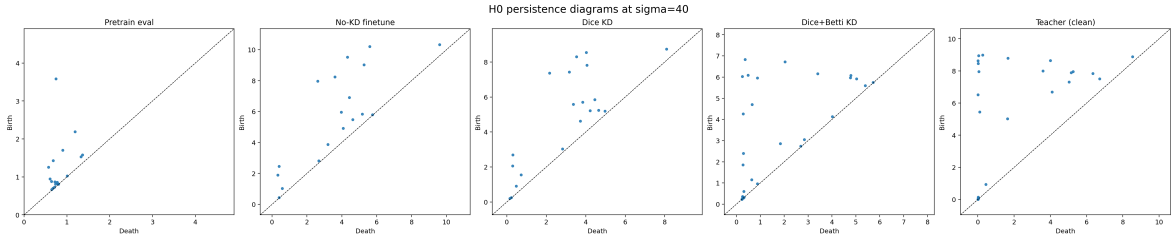

Figure 5. Persistence diagrams at $\sigma = 40$ for the compared methods, providing a birth–death view of topological events in the reconstructed response field. The Dice+Betti student model shifts its diagram toward the teacher model diagram, indicating closer agreement in the birth and merging behavior of objectness components.

The visualization explains why direct noisy evaluation is fragile. As noise increases, the objectness field becomes less structured. Responses may weaken, background activations may become more visible, and nearby object responses may be harder to separate.

The comparison at $\sigma = 40$ further shows that the proposed distillation changes both the spatial organization of the reconstructed response field and its topological persistence structure.

In particular, the Dice+Betti student model is visually closer to the teacher model in the heatmap overlay. It is also closer in the barcode and persistence diagram. This directly reflects the objective of aligning H_0 persistence between $\widehat{H}^{\text{stu}}$ and $\widehat{H}^{\text{tea}}$.

These qualitative results agree with the quantitative collapse of the pretrain-eval baseline under strong corruption. They also support the motivation for topology-aware distillation: the student model is encouraged to preserve the teacher model's peak–valley organization, not only local activation magnitudes.

5.6 Component Analysis

The comparison between Dice KD and Dice+Betti KD gives a compact component analysis of the proposed objective. Dice supervision alone aligns the reconstructed heatmaps by volume. It does not directly control whether object responses stay separated, nor whether noisy bridges merge neighboring peaks.

Adding the Betti term improves mAP stability under medium and high noise. This suggests that topology-aware supervision is most useful as a structural regularizer on top of

local heatmap alignment.

At the same time, the gains are not uniform across all metrics. Some precision and recall entries remain close to, or slightly below, the baseline. Thus, the topological term improves the global organization of objectness responses, but it does not automatically optimize every detection statistic by itself.

5.7 Analysis

The results suggest that the main benefit of the proposed method is structural preservation under noise. Standard heatmap alignment encourages similar activation magnitudes. It does not directly constrain whether peaks stay separated, whether valleys stay low, or whether nearby responses merge at the correct threshold levels.

The topology-aware term addresses this gap by regularizing the birth and merging behavior of connected components in the reconstructed field.

This also explains why the hybrid objective works better than pure topology supervision. Topological consistency captures global structural behavior, but it can miss local differences that still matter for precise detection. By combining topology-aware matching with Dice-based alignment, the model preserves both global connectivity and local geometric support. This leads to better overall robustness.

The method also inherits a limitation of teacher model–student model distillation. The topological target comes from the teacher model heatmap. Therefore, the student model is encouraged to reproduce the teacher model's structure, including possible teacher errors. The effectiveness of topology-aware distillation depends on the reliability of the teacher model prediction. In very crowded scenes, or when objects are weakly separated, the reconstructed field may contain ambiguous peak interactions. This can make the persistence structure less stable.

The experiments results indicate that topology-aware distillation provides a meaningful structural prior for robust object detection under noisy inputs.

6. Discussion

The results show that topology-aware distillation is most useful under stronger corruption, while its gains at low noise remain limited. The following discussion summarizes its cost, mechanism, limitations, and possible extensions.

6.1 Computational Complexity

The proposed method restricts topological regularization to 0-dimensional persistent homology (H_0). In this case, features correspond to connected components in the superlevel filtration. On a 2D grid, H_0 persistence can be computed efficiently with a Union-Find scheme. The algorithm activates pixels from high to low values and merges neighboring components. This gives an amortized complexity of $O(n\,\alpha(n))$, where n is the number of grid locations and $\alpha(\cdot)$ is the inverse Ackermann function.

The main overhead comes from ordering pixel intensities before component merging. In practice, the reconstructed heatmap is discretized into a finite set of levels, which makes the ordering step easier to implement.

The method also does not change the detector architecture. The topology term is used only during training-time supervision, and inference still uses the original detector without any extra topological module. This makes the approach more practical than higher-order topological constraints, whose cost grows quickly on dense objectness maps.

However, the current implementation still has a high training-time cost. In our experiments, Dice+Betti distillation takes about 2 hours per epoch. Pure Dice distillation takes about 1.5 minutes per epoch, and No-KD fine-tuning takes about 40 seconds per epoch.

The hybrid objective improves robustness under medium and high noise, but the gain is modest compared with the extra cost. From a practical view, the current Dice+Betti implementation has low cost-effectiveness unless the application specifically needs stronger structural preservation of the objectness landscape.

6.2 Why Topology Helps

The proposed loss can be seen as structural distillation on the objectness head. The teacher model heatmap $\widehat{H}^{\text{tea}}$ and the student model heatmap $\widehat{H}^{\text{stu}}$ are reconstructed on the same stride-8 grid. They can therefore be compared directly as grayscale fields.

Under corruption, the most harmful changes are often structural rather than purely pixelwise. A true response may split into several nearby peaks. Two adjacent responses may merge through a raised bridge. Spurious peaks may also appear in the background. By aligning the H_0 persistence diagrams of $\widehat{H}^{\text{stu}}$ and $\widehat{H}^{\text{tea}}$, the loss constrains the birth and merging behavior of connected components. This behavior is not directly controlled by ordinary local alignment.

This interpretation also explains the experimental behavior. Dice alone already gives a strong volumetric alignment signal. Its gains are less stable under severe corruption because it does not directly preserve connectivity structure. The hybrid Dice+Betti objective works better because it combines local similarity with structural regularization.

At the same time, the topology term should not be treated as a complete replacement for volumetric alignment. Persistent homology captures global connectivity behavior, but it is less sensitive to some local geometric differences that still matter for precise localization and confidence calibration. For this reason, the best results come from using topology-aware supervision as a complement to conventional heatmap matching.

6.3 Limited Improvement

Another limitation is that topological similarity of reconstructed heatmaps does not directly guarantee better final separation metrics. The persistence barcode and diagram can show that the student heatmap becomes more similar to the teacher heatmap.

Final detections, however, are still produced after confidence filtering and NMS. This post-processing step may suppress, merge, or reorder candidate boxes. It can weaken the link between heatmap-level topology and final box-level separation performance.

As a result, better topological alignment should be interpreted as evidence of improved

structural consistency, not as a guarantee of large gains in all detection metrics.

6.4 Future Work

Several extensions are worth exploring. A natural next step is to add multi-resolution constraints to the current fused-field formulation. The present method aggregates decoded predictions from multiple detection scales into a shared stride-8 heatmap. It does not directly regularize the intermediate per-level fields before fusion. Future work may combine the current unified topology loss with auxiliary cross-scale constraints. These constraints could encourage coherent persistence behavior across both the shared supervisory field and the native detector resolutions.

It would also be valuable to study adaptive weighting between Dice and topology terms. A fixed balance may not be ideal for all training stages or noise regimes. An adaptive design may let the model rely more on local geometric alignment in easy settings, and more on structural regularization when corruption becomes severe. Finally, broader evaluation on multi-class benchmarks and more crowded real-world scenes would show how well topology-aware distillation generalizes beyond the current proof-of-concept setting.

7. Conclusion

This work explored topology-aware knowledge distillation for improving the noise robustness of YOLO11 object detection.

Instead of matching only final bounding boxes or local response values, we reconstructed a unified stride-8 objectness landscape from pre-NMS decoded predictions. Persistent homology was then used to compare the structural behavior of teacher model and student model heatmaps.

The proposed Dice+Betti objective ensures the student model to preserve the teacher model's peak–valley organization under noisy inputs. The qualitative visualization of heatmaps, persistence barcodes, and persistence diagrams shows that the distilled student model can become topologically closer to the clean teacher model.

The experiments indicate that this structural supervision is most useful under medium

and high noise. In these settings, objectness responses are more likely to fragment, merge, or produce spurious peaks.

At the same time, the overall performance improvement is limited and not uniform across all detection metrics. The current Dice+Betti implementation also has a high training-time cost compared with pure Dice distillation and No-KD fine-tuning.

Therefore, the method is best viewed as a validation of the concept. It shows that topology can provide meaningful structural guidance for robust detection. Further work is still needed to reduce computational cost and strengthen the link between heatmap-level topological consistency and final box-level detection performance.

8. Code Availability

The code used in this thesis is available at `https://github.com/cz024/ug_proj`.

9. Acknowledgements

The author thanks Professor Hongwei Lin for inspiring the research topic of this thesis. The author also thanks Professor Yifei Zhu for helpful discussions and guidance on the experimental ideas. The author further thanks Bohan Shen, for introducing AI tools for manuscript formatting. ChatGPT (OpenAI) was used to assist with language refinement and structural editing of the manuscript. All mathematical content, statements, and references in the manuscript were checked and approved by the author. The author takes full responsibility for the final text.

A Formula Summary

Table 5. Summary of formulas and symbols used in this work.

Item	Formula	Meaning
Input space	$\mathcal{X} := \{1, \dots, \text{Color}\}^{\text{Height} \times \text{Width}}$	Image domain.
Detection output space	$\mathcal{Y} := \{S \subset \mathbb{R}^4 \times \mathcal{C} \times [0,1] \mid \lvert S\rvert < \infty\}$	Finite detection sets.
Detector mapping	$f : \mathcal{X} \to \mathcal{Y}, \quad f(x) := \{(b_i, c_i, s_i)\}_{i=1}^{N(x)}$	Set-valued output.
Box coordinates	$b_i := (x_i^{(1)}, y_i^{(1)}, x_i^{(2)}, y_i^{(2)}) \in \mathbb{R}^4$	Image-coordinate box.
Multi-scale index set	$\mathcal{I} := \bigsqcup_{s \in \{8,16,32\}} \mathcal{I}_s$	All decoded candidates.
Candidate score and class	$s_i := \max_r \sigma(z_{i,r}), \quad c_i := \arg\max_r \sigma(z_{i,r})$	Confidence and class.
Retained detections	$\mathcal{R}(x) \subseteq \mathcal{I}, \quad N(x) := \lvert\mathcal{R}(x)\rvert$	Indices kept after filtering and NMS.
Shared grid	$\Omega_8 := \{1, \dots, H_8\} \times \{1, \dots, W_8\}$	Stride-8 lattice.
Grid size	$H_8 := \lceil \text{Height}/8 \rceil, \quad W_8 := \lceil \text{Width}/8 \rceil$	Spatial size of Ω_8.
Reconstructed heatmap	$H(u,v) := \sum_{i \in \mathcal{I}} s_i \, K((u,v); b_i)$	Buildup from pre-NMS boxes.
B-spline kernel	$K_{\text{bspline}}((u,v); b_i) := \lambda \beta_3\left(\frac{v - m_i^x}{\rho_i^x}\right) \beta_3\left(\frac{u - m_i^y}{\rho_i^y}\right)$	Box-induced kernel.
Compressed heatmap	$\widehat{H}(u,v) := 1 - \exp(-H(u,v))$	Bounded heatmap on Ω_8.
Reconstruction operator	$\Phi : \{(b_i, s_i)\}_{i \in \mathcal{I}} \mapsto \widehat{H}$	Weighted boxes to heatmap.
Superlevel set	$\Omega_\tau := \{(u,v) \in \Omega_8 \mid \text{quant}_{256}(\widehat{H}(u,v)) \geq \tau\}$	High-response region.
Persistence diagram	$\text{Dgm}(\widehat{H}) := \{(b_\tau, d_\tau) \mid [b_\tau, d_\tau) \in \mathcal{B}_\tau\} \subset \mathbb{R}^2$	Birth–death pairs.
Teacher/student heatmaps	$\widehat{H}^{\text{tea}} := \Phi(P^{\text{tea}}), \quad \widehat{H}^{\text{stu}} := \Phi(P^{\text{stu}})$	Fields compared in KD.
Dice loss	$\mathcal{L}_{\text{dice}} := 1 - \frac{2 \sum_{\Omega_8} \widehat{H}^{\text{stu}} \widehat{H}^{\text{tea}}}{\sum_{\Omega_8} \widehat{H}^{\text{stu}} + \sum_{\Omega_8} \widehat{H}^{\text{tea}}}$	Betti-matching code call: `DiceLoss(sigmoid=True)`.
Topology loss	$\mathcal{L}_{\text{topo}} := \sum_{q \in \text{Dgm}_0(\widehat{H}^{\text{stu}})} 2 \lVert q - \tau(\widehat{H}^{\text{stu}}, \widehat{H}^{\text{tea}})(q) \rVert_2^2$	Matches H_0 persistence.
KD objective	$\mathcal{L}_{\text{kd}} := \mathcal{L}_{\text{dice}} + \alpha \mathcal{L}_{\text{topo}}$	Hybrid heatmap loss.
Training objective	$\mathcal{L}_{\text{train}} := \mathcal{L}_{\text{det}} + \lambda_{\text{kd}} \mathcal{L}_{\text{kd}}$	Detection plus distillation.
IoU metric	$\text{IoU}(B_p, B_{gt}) := \frac{\lvert B_p \cap B_{gt}\rvert}{\lvert B_p \cup B_{gt}\rvert}$	Box overlap.
Precision/Recall	$\text{Precision} := \frac{\text{TP}}{\text{TP+FP}}, \quad \text{Recall} := \frac{\text{TP}}{\text{TP+FN}}$	Detection accuracy metrics.
Average Precision	$\text{AP} := \int_0^1 \text{Precision}(r)\, dr$	Area under PR curve.
mAP@50	$\text{mAP@50} := \frac{1}{C} \sum_{c=1}^{C} \text{AP}_c^{(\text{IoU} \geq 0.5)}$	AP at IoU 0.5.
mAP@0.5:0.95	$\text{mAP@0.5:0.95} := \frac{1}{T} \sum_{t \in \mathcal{T}} \text{mAP@}t$	AP averaged over IoU thresholds.

References

[1] J. R. Clough, N. Byrne, I. Oksuz, V. A. Zimmer, J. A. Schnabel, and A. P. King. A topological loss function for deep-learning based image segmentation using persistent homology. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8766–8778, 2022. DOI: 10.1109/TPAMI.2020.3013679.

[2] X. Dai, Z. Jiang, Z. Wu, Y. Bao, Z. Wang, S. Liu, and E. Zhou. General instance distillation for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7842–7851, 2021.

[3] H. Edelsbrunner and J. Harer. *Computational Topology: An Introduction*. American Mathematical Society, 2010.

[4] J. Kim, J. You, D. Lee, H. Y. Kim, and J.-H. Jung. Do topological characteristics help in knowledge distillation? In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235, pages 24674–24693. PMLR, 2024.

[5] C. Michaelis, B. Mitzkus, R. Geirhos, E. Rusak, O. Bringmann, A. S. Ecker, M. Bethge, and W. Brendel. Benchmarking robustness in object detection: autonomous driving when winter is coming. In *NeurIPS Workshop on Machine Learning for Autonomous Driving*, 2019.

[6] M. J. Mirza, C. Buerkle, J. Jarquin, M. Opitz, F. Oboril, K.-U. Scholl, and H. Bischof. Robustness of object detectors in degrading weather conditions. In *2021 IEEE 24th International Conference on Intelligent Transportation Systems (ITSC)*, 2021.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[8] S. Stucki, J. A. F. de Sousa, and G. R. G. Lanckriet. Topologically faithful image segmentation via induced matching of persistence barcodes. In *International Conference on Machine Learning (ICML)*, 2023.

[9] Ultralytics. Yolov11: ultralytics yolo. `https://github.com/ultralytics/ultralytics`, 2024.

[10] T. Wang, L. Yuan, X. Zhang, and J. Feng. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4933–4942, 2019.

[11] Z. Yang, Z. Li, X. Jiang, Y. Gong, Z. Yuan, D. Zhao, and C. Yuan. Focal and global knowledge distillation for detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[12] Z. Zheng, R. Ye, P. Wang, D. Ren, W. Zuo, Q. Hou, and M.-M. Cheng. Localization distillation for dense object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.