TUM SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY – CONNECTED MOBILITY

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

Test-Time Accuracy Indicators for Object Detection

Zhenghao Lu

TUM SCHOOL OF COMPUTATION,
INFORMATION AND TECHNOLOGY –
CONNECTED MOBILITY

TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelor's Thesis in Informatics

Test-Time Accuracy Indicators for Object Detection

Testzeit-Genauigkeitsindikatoren für die Objekterkennung

Author:	Zhenghao Lu
Examiner:	Prof. Dr.-Ing. Jörg Ott
Supervisor:	Wei Geng, M.Phil., M.Eng
Submission Date:	30.06.2026

I confirm that this bachelor's thesis in informatics is my own work and I have documented all sources and material used.

Munich, 30.06.2026 Zhenghao Lu

Acknowledgments

I would like to thank my supervisor, Wei Geng, for his patience and support throughout my thesis. His suggestions and feedback especially helped me develop better academic habits and gave me a first introduction to this research field.

I would also like to thank my family and friends for their continuous support and encouragement during my studies and the completion of this thesis. Although I still have much to learn, this journey has allowed me to gain knowledge and a clearer understanding of the path ahead.

Abstract

Video object detection systems usually have to balance detection accuracy and computational efficiency. Powerful detectors can provide more reliable prediction results but come with higher inference costs, while lightweight detectors are more efficient but may not perform well on difficult frames. This thesis investigates whether frame-level difficulty can be estimated during testing and used to adaptively route between weak detectors and strong detectors.

One important characteristic of video frames is temporal consistency, which indicates that detection results in consecutive frames should also remain consistent. Existing quality estimation methods often require additional learning components or fail to explicitly utilize temporal consistency in videos. To address this issue, this thesis proposes a temporal-consistency-based indicator called PTC-IoU that utilizes detection, association and localization consistency and estimates frame difficulty by analyzing the stability of detector outputs between adjacent frames, without requiring additional training.

This method was evaluated on the MOT17 dataset, using the weak detector Faster R-CNN MobileNetV3 and the strong detector Faster R-CNN ResNet50-FPN-V2. The evaluation was conducted from three aspects: ranking quality, routing effectiveness and computational performance. The results show that PTC-IoU provides meaningful frame-level difficulty estimation. Compared with learning-based baselines such as GFL and IoU-Net, its routing performance is competitive and it only introduces minor additional runtime overhead. Overall, the results indicate that temporal consistency can serve as a lightweight and practical signal in adaptive video object detection.

Contents

1. Introduction

1.1. Background

Object detection is one of the core tasks in modern computer vision, which aims to classify and localize specific objects in images [1]. In recent years, progress in deep learning has significantly improved the performance of object detection. Modern object detectors, including two-stage detectors such as Faster R-CNN [2] and more recent Transformer-based detectors [3], have shown excellent performance on benchmark datasets. Therefore, they are widely used in autonomous driving, traffic monitoring and robotics.

In practical applications, images rarely appear alone. On the contrary, video-based scenarios are more common, including continuous and temporally related frames [4]. As illustrated in Figure 1.1, such scenarios often require real-time processing under limited computational resources. This creates a trade-off between detection accuracy and computational efficiency.

At the same time, the temporal structure of videos provides additional information that is not available in isolated images. Since objects usually move continuously across adjacent frames, stable detection results over time may indicate that frames in this video are relatively easy to process, while inconsistent detections, ambiguous associations or unstable localization may reflect higher frame-level difficulty. Therefore, temporal consistency can be used as a proxy signal for estimating frame difficulty and supporting adaptive detector selection.

1.2. Motivation

Over the past decade, object detection has been widely studied and many methods have been proposed to improve its performance. Recent progress has benefited from deep learning-based models, such as Convolutional Neural Network detectors like Faster R-CNN [2], while Transformer-based detection models have further improved detection performance [3]. Most of these methods focus on designing stronger network architectures, adopting more advanced feature representations or expanding the size of training data.

Although these methods do achieve significant performance improvements, they tend to rely on increasingly complex model designs and higher computational cost [2, 5, 6, 7]. In contrast, less attention is paid to the stability analysis of detection results, especially in video sequences, where temporal consistency is particularly important. In addition, many practical applications have real-time requirements. For example, in autonomous driving, the perception system must efficiently process the received image frames to ensure timely decision-making [8]. Therefore, detection accuracy alone is not enough and efficiency is also very important.

In this trade-off, a key challenge is to determine when the output of a lightweight detector is reliable enough and when a stronger detector is needed. In video object detection, this judgment can be assisted by temporal consistency: if the detection is reliable, the detection results are expected to remain relatively continuous and stable across adjacent frames. Conversely, inconsistent detection results may indicate that the current frame is less reliable and therefore more difficult to process. Therefore, this thesis aims to estimate frame-level difficulty based on temporal consistency during the testing stage. Subsequently, the estimated difficulty can be used as a routing signal to adaptively select detectors [9, 10].

1.3. Contributions

To address the trade-off between detection accuracy and computational efficiency, this thesis studies frame-level difficulty estimation for adaptive video object detection. The core idea is to judge whether a frame can be reliably processed by a lightweight detector or should be routed to a more powerful detector for processing.

The main contributions of this thesis are summarized as follows.

- This thesis proposes a training-free frame-level difficulty estimation metric called PTC-IoU for video object detection. This method does not rely on additional learned prediction heads; instead, it estimates difficulty by leveraging the temporal consistency of detector outputs between adjacent frames. To capture this temporal consistency from multiple perspectives, PTC-IoU combines three sub-dimensions: detection consistency, association consistency and localization consistency. These sub-dimensions measure the temporal stability of detection results, the clarity of object associations and whether the bounding boxes follow a temporally consistent motion pattern respectively.
- The proposed difficulty estimator was evaluated from multiple perspectives, including ranking quality, routing effectiveness and computational performance. Ranking metrics were used to analyze whether the estimated difficulty reflected frame-level detection performance; routing experiments were conducted to verify whether these scores could support adaptive detector allocation; and runtime analysis was used to measure the additional overhead introduced by PTC-IoU. The results show that, compared with the weak detector, PTC-IoU not only provides useful difficulty signals but also introduces only limited additional overhead.

1.4. Challenges

It is definitely not easy to estimate frame-level difficulty for object detection and it involves several challenges.

- First, frame-level difficulty cannot be directly measured during inference due to the absence of ground-truth annotations at test time. Therefore, the difficulty of a frame

has to be inferred from other observable signals produced by the detector itself, such as temporal consistency in video object detection.

- Second, the difficulty indicator should be computationally lightweight. If difficulty estimation introduces excessive additional overhead, the benefit of adaptive routing may be reduced or even lost. This motivates a training-free design that does not require additional learned prediction heads.

- Third, temporal consistency does not always directly correspond to detection quality. Although stable detection results across adjacent frames often indicate easier frames, this assumption may fail in some cases. One typical case is consistently incorrect detection. For example, if an object is missed by the detector over several consecutive frames, the detection output may still appear temporally stable because the same error is repeated over time. In this case, PTC-IoU may regard the frame as easy, while the actual detection quality can be low.

  Another case is correct but temporally unstable detection. For example, due to fast object motion, motion blur, occlusion, camera movement, or detector uncertainty, correct bounding boxes may fluctuate across adjacent frames. In this case, temporal consistency may become low even though the detector still produces reasonable predictions.

  Figure 1.2 illustrates these two cases. In the first example, the detector does not consistently detect the vehicle that is entering the main road. As a result, the repeated missing detection may still produce temporally stable outputs, even though the actual detection quality is low. The second example shows a fast-moving vehicle with strong motion blur, where large appearance and localization changes may reduce temporal consistency even though the object is still visually present. Therefore, temporal consistency should be regarded as a proxy signal rather than a perfect difficulty indicator, and the relationship between temporal consistency and frame difficulty needs to be carefully evaluated.

- Finally, ranking quality does not necessarily translate into routing gain. Even if a frame is correctly identified as difficult, reprocessing it with a stronger detector may not always bring significant improvement. Therefore, both ranking-based and routing-based evaluations are needed.

1.5. Thesis Structure

The rest of this thesis is structured as follows.

Section 2 reviews related work on object detection and difficulty analysis.

Section 3 introduces the proposed method for frame-level difficulty estimation in video sequences.

Section 4 describes the experimental setup and datasets used in this study.

Section 5 presents the results and evaluation of the proposed method.

Finally, Section 6 concludes the thesis and discusses potential limitations and future work.

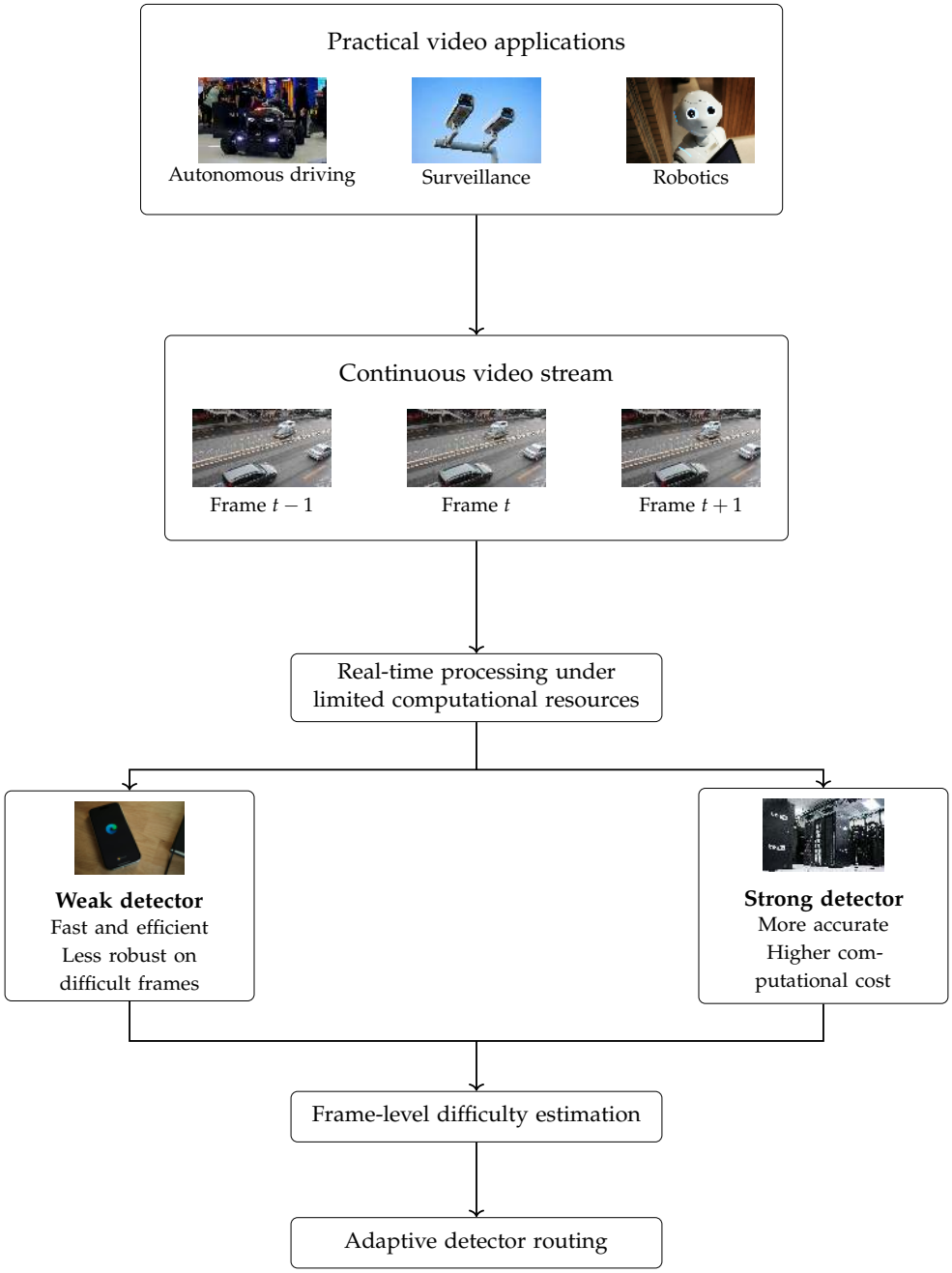
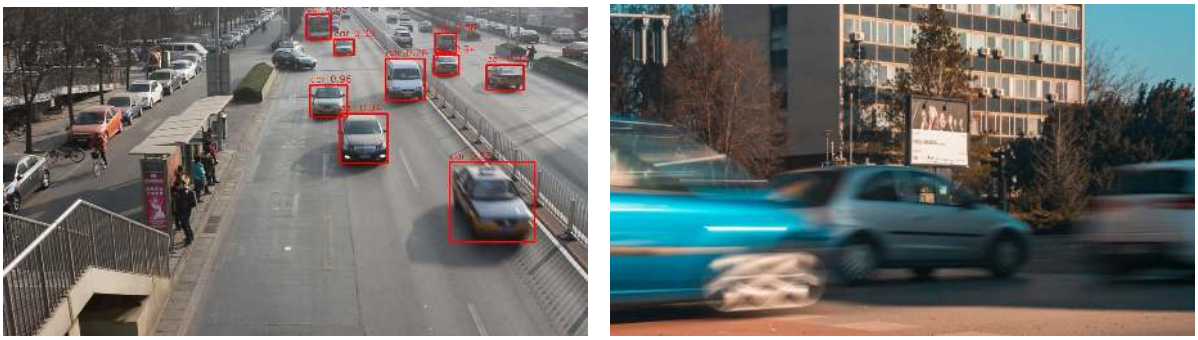


Figure 1.1.: Conceptual motivation of adaptive video object detection in practical applications.



(a) Consistently incorrect detection (b) Correct but temporally unstable detection

Figure 1.2.: Examples of two failure cases of temporal consistency: consistently incorrect detection and correct but temporally unstable detection.

2. Related Work

2.1. Temporal Consistency in Video Analysis

Unlike static images, video data have obvious temporal continuity and adjacent frames are highly related in appearance and object motion. The use of this temporal consistency has been proven to be beneficial for a variety of video understanding tasks, including object detection, segmentation and tracking. In particular, temporal consistency has become a key factor in improving the stability and robustness of model predictions over time, which is crucial for achieving reliable performance in video scenarios.

Recent studies have clearly incorporated temporal consistency into the learning objectives. For example, video shadow detection methods based on spatio-temporal interpolation consistency training, such as Video Shadow Detection via Spatio-Temporal Interpolation Consistency Training [11], use spatio-temporal interpolation and alignment strategies to promote the consistency of predictions between different frames. These methods encourage the model to generate temporally stable outputs, thus reducing flickering predictions and improving overall video-level performance.

In addition to shadow detection, similar ideas have also been applied to the field of video object detection and tracking. For example, Deep Feature Flow [4] takes advantage of temporal redundancy in videos to propagate deep features across frames, so as to reduce duplicate computation while maintaining recognition performance. These studies show that temporal information is not only an inherent attribute of video data, but can also be actively used to improve the efficiency and stability of video analysis systems.

However, although temporal consistency has been widely studied in modeling and training, it is rarely used as a lightweight test-time signal to estimate frame-level detection difficulty. Many existing methods use temporal information to improve the prediction results of the model, but do not clearly provide a simple frame-level proxy indicator to determine whether the current detection results are reliable. This motivates us to use temporal consistency not only as a training objective, but also as a test-time indicator in adaptive video object detection.

2.2. Detection Quality Estimation

In addition to standard object detection models, some methods have also been proposed to evaluate the quality or reliability of detection results. The starting point of these methods is the observation that classification confidence alone does not always accurately reflect localization quality or detection reliability. A detection result may have a high classification score, but still be poorly localized, which can reduce the effectiveness of post-processing and

ranking.

IoU-Net [12] predicts the localization quality of detected bounding boxes by estimating the Intersection-over-Union with the corresponding ground-truth boxes. The predicted localization confidence can be used to optimize detection ranking and refine the selection of high-quality bounding boxes. In this way, IoU-Net introduces an additional learned quality estimation component to more effectively distinguish between well-localized detection results and poorly localized results.

Generalized Focal Loss (GFL) [13] further integrates classification confidence and localization quality into a unified representation. Unlike previous methods that treat classification and localization quality as independent factors, GFL encourages predicted scores to better reflect the overall quality of detections, so that detection scores are more informative for ranking and post-processing.

These methods show that estimating detection quality is important for improving the reliability of object detection systems. However, they usually require additional learned prediction heads and extra inference operations, such as RoI feature extraction or quality-head prediction. In contrast, the proposed PTC-IoU indicator estimates frame-level difficulty through temporal consistency without training additional quality prediction heads. Therefore, it provides a lightweight alternative for test-time difficulty estimation in video object detection.

Table 2.1 compares these methods from multiple perspectives, including input format, use of temporal information, training requirements, output level, computational characteristics, and main limitations. This comparison highlights that IoU-Net and GFL estimate quality through learned components based on single-image detector features, while the proposed PTC-IoU indicator estimates frame-level difficulty from temporal consistency between adjacent-frame detection outputs without additional training.

2.3. Adaptive Inference and Routing

Adaptive inference aims to improve the trade-off between accuracy and computational cost by allocating computational resources according to the difficulty or importance of each input. Instead of using the same model for all samples, it can use a lightweight model to process simpler samples and hand over more difficult samples to a stronger model or additional processing stages.

Relevant ideas have been studied in the framework of selective classification and learning-to-defer. Selective classification allows the model to avoid making predictions when the confidence is low [14, 15]. Learning-to-defer methods further explore how the model decides whether to make predictions by itself or defer the decision to other experts [16, 17]. These methods show that adaptive decision-making can improve reliability by adopting different processing methods for different inputs.

In object detection, adaptive routing and edge offloading techniques are also explored to reduce computational costs while maintaining detection performance. For example, edge-cloud collaborative object detection uses a difficult-case discriminator to decide whether the input should be processed locally or offloaded to a more powerful cloud model [9]. Selective

Table 2.1.: Comparison between learned detection quality estimation methods and the proposed temporal-consistency-based indicator.

Comparison Dimension	IoU-Net	GFL	PTC-IoU (Ours)
Core idea	Predict localization IoU	Joint classification-localization quality	Measure temporal consistency
Input format	Single-image detections and RoI features	Single-image detector features	Adjacent-frame detection outputs
Temporal information	Not explicitly used	Not explicitly used	Explicitly used
Requires additional training	Yes	Yes	No
Output level	Box-level quality	Box-level quality	Frame-level difficulty indicator
Main computation	GPU-based feature and head inference	GPU-based quality-head inference	Lightweight post-processing on detections
Test-time overhead	Additional RoI and head computation	Additional quality-head computation	Matching and score aggregation
Main limitation	Requires learned quality prediction	Requires modified training objective	Depends on the reliability of temporal consistency

offloading has also been studied from the perspective of budget-adaptive routing, where the system decides whether to skip the weak model and directly use the strong model under a given computational budget [18].

Similarly, edge offloading methods study how to assign object detection tasks to devices with different computational capacities [10]. Recent work on scalable multitask offloading further explores backbone sharing to reduce redundant computation across multiple tasks in edge-cloud systems [19].

Compared with these methods, this thesis focuses on frame-level difficulty estimation in video object detection. The goal is not to train an independent routing network, but to derive a lightweight test-time difficulty signal from temporal consistency. This signal can be used to support adaptive routing between weak detectors and strong detectors.

2.4. Evaluation Metrics for Object Detection and Tracking

Evaluation metrics are the main way to measure the performance of object detection and multi-object tracking systems. In object detection, the most commonly used metric is mean Average Precision (mAP), which evaluates the balance between precision and recall under different confidence thresholds [1]. Variants such as 11-point interpolated average precision and COCO-style mAP further refine this evaluation by taking different IoU thresholds into account. In addition to mAP, precision and recall are also often reported to provide complementary information about detection performance.

Although these metrics are usually reported at the dataset level, their calculations are based on the spatial matching between predicted detection results and ground-truth objects in a single image or frame. Therefore, they mainly reflect frame-level spatial detection quality and do not explicitly model the temporal relationship between adjacent frames. This limitation is especially important in video scenarios, because temporally inconsistent detection results, such as missed detections in some frames or unstable object localization, may still obtain acceptable frame-level detection scores.

In order to consider the temporal factor, the multi-object tracking field introduces sequence-level evaluation metrics that go beyond single-frame detection quality. Traditional tracking metrics such as Multiple Object Tracking Accuracy (MOTA) and Multiple Object Tracking Precision (MOTP) [20] summarize false positives, missed detections and localization-related errors. Identity-based metrics, such as IDF1 [21], further evaluate whether objects can maintain consistent identities across different frames.

Despite these advances, existing evaluation metrics are still limited for test-time difficulty estimation. Detection-based metrics mainly focus on spatial accuracy, while tracking metrics are usually designed for sequence-level post-evaluation based on ground-truth annotations. Therefore, they are not directly applicable to lightweight frame-level quality estimation during inference.

2.5. Higher-order Tracking Metrics

In order to overcome the limitations of traditional tracking metrics, recent research has proposed higher-order evaluation frameworks, which can comprehensively capture multiple aspects of tracking performance. A representative example is HOTA, which provides a unified evaluation method for detection, association and localization quality in multi-object tracking [22].

Specifically, HOTA decomposes tracking performance into three complementary components: detection accuracy, which is used to measure how well objects are detected; association accuracy, which is used to evaluate the consistency of object identities across different frames; and localization accuracy, which is used to capture the spatial alignment between predicted results and ground-truth objects. By integrating these components, HOTA provides a more balanced and interpretable evaluation method compared with earlier metrics such as MOTA or IDF1.

The introduction of HOTA highlights the importance of considering both spatial and temporal factors in the quality evaluation of video predictions. In particular, the explicit inclusion of association quality reflects the key role of temporal consistency in video prediction and evaluation. This is closely related to the research motivation of this thesis, in which detection, association and localization consistency are also regarded as complementary aspects of frame-level difficulty.

However, despite its advantages, HOTA is mainly designed as a comprehensive offline evaluation metric that requires access to ground-truth annotations of the entire sequence. Its formulation focuses on global sequence-level evaluation and does not directly provide

a lightweight frame-level proxy indicator that can be efficiently calculated at test time. In contrast, the proposed PTC-IoU indicator adopts similar conceptual dimensions, but these dimensions are estimated through detector outputs from adjacent frames and there is no need to rely on ground-truth annotations during inference.

These limitations show that although higher-order metrics such as HOTA can provide valuable insights into tracking performance, it is still necessary to develop alternative formulations that can maintain interpretability while enabling efficient, fine-grained and test-time estimation of detection difficulty in video scenarios.

3. Method

3.1. Overview

In this section, the proposed Proxy-based Temporal Consistency IoU (PTC-IoU) indicator and its mathematical formulation are introduced. Following the decomposition concept of HOTA [22], this indicator is divided into three interpretable sub-dimensions: detection consistency, association consistency and localization consistency. Different aspects of temporal consistency may reflect different sources of frame processing difficulties.

Since the proposed method operates in a post-model manner, frame-level detection results are first generated by weak and strong detectors on the MOT17 dataset [23]. Then, the proposed temporal consistency method is applied to the detector outputs to calculate the corresponding difficulty scores.

Finally, this proposed method is evaluated from two perspectives: ranking-oriented and routing-oriented. Through comparisons with multiple baselines, these experiments aim to explore whether temporal consistency can provide meaningful signals for frame-level difficulty ranking and adaptive routing in video object detection.

3.2. Temporal Consistency Formulation

Given two adjacent video frames at time step $t-1$ and t, the corresponding detection sets are denoted as:

$$D_{t-1} = \{d_i^{t-1}\}_{i=1}^{N_{t-1}}$$

and

$$D_t = \{d_j^t\}_{j=1}^{N_t}$$

where N_{t-1} and N_t represent the numbers of detections in the previous frame and current frame, respectively.

Each detection is represented as:

$$d = (b, s, c)$$

where b denotes the bounding box, s denotes the confidence score and c denotes the object category.

To measure temporal consistency between adjacent frames, detections are matched using class-aware Intersection-over-Union (IoU) [1]. For a detection pair (d_i^{t-1}, d_j^t), the IoU value is defined as:

$$\mathrm{IoU}(d_i^{t-1}, d_j^t) = \frac{\mathrm{Area}(b_i^{t-1} \cap b_j^t)}{\mathrm{Area}(b_i^{t-1} \cup b_j^t)}$$

For each detection in the current frame, the best matched detection in the previous frame is determined by:

$$m(j) = \arg\max_i \mathrm{IoU}(d_i^{t-1}, d_j^t)$$

where only detections belonging to the same category are considered during the matching process.

The proposed method is based on the observation that **relatively easy video frames should produce temporally continuous and stable bounding boxes across adjacent frames**. When an object is consistently detected in adjacent frames with stable localization and clear matching relationships, the corresponding frame is more likely to be reliably processed by the detector. In contrast, unstable detections, ambiguous associations or inconsistent localization may indicate a higher frame-level difficulty.

Figure 3.1 provides an example of this observation. In this relatively simple scene, the number of visible objects is limited and the detector produces highly consistent results across two adjacent frames. Most detected vehicles appear in both frames and their bounding boxes remain stable in terms of spatial location and confidence. This supports the assumption that temporally continuous detection outputs can serve as an indicator of lower frame-level difficulty.

(a) Frame $t-1$ (b) Frame t

Figure 3.1.: Example of temporally stable detection results in two adjacent frames. In this relatively simple scene with a limited number of objects, detections remain consistent in both object presence and bounding-box localization, indicating lower frame-level difficulty.

Based on this observation, temporal consistency is decomposed into three interpretable sub-dimensions [22]:

- Detection Consistency (PTC-Det)
- Association Consistency (PTC-Ass)
- Localization Consistency (PTC-Loc)

These three consistency sub-dimensions are described in the corresponding sections.

3.3. Difficulty Score Construction

3.3.1. Detection Consistency (PTC-Det)

Detection consistency is used to measure whether an object can remain consistently detected in adjacent video frames. In difficult frames, due to occlusion, motion blur or crowded scenes, the detection results may disappear, reappear or fluctuate [22]. Therefore, unstable detection can be regarded as an indicator of higher frame-level difficulty.

Let $\mathcal{M}_t$ denote the set of one-to-one matched detection pairs between frame $t-1$ and frame t obtained by class-aware IoU matching. For each matched pair $(i,j) \in \mathcal{M}_t$, the confidence contribution of the matched pair is defined as the smaller confidence score of the two detections:

$$w_{ij} = \min(s_i^{t-1}, s_j^t).$$

The confidence-weighted matched detection score is then defined as:

$$W_t^{\text{match}} = \sum_{(i,j)\in\mathcal{M}_t} w_{ij}.$$

The total confidence scores of the detection sets in two adjacent frames are:

$$W_{t-1} = \sum_{i=1}^{N_{t-1}} s_i^{t-1}, \qquad W_t = \sum_{j=1}^{N_t} s_j^t.$$

The basic intuition of detection consistency can first be formulated using a count-based Jaccard similarity. Let $|\mathcal{M}_t|$ denote the number of matched detections between two adjacent frames. A simple consistency score can then be defined as:

$$\text{PTC-Det}_t^{\text{count}} = \frac{|\mathcal{M}_t|}{N_{t-1} + N_t - |\mathcal{M}_t|}.$$

This formulation measures the overlap ratio between the detection sets in adjacent frames. However, directly counting detections treats all detections equally and does not consider the reliability of different predictions. In practice, low-confidence detections are usually less stable and more likely to introduce noise into temporal consistency estimation.

Based on these terms, the proposed detection consistency score is defined as a confidence-weighted Jaccard similarity:

$$\text{PTC-Det}_t = \frac{W_t^{\text{match}}}{W_{t-1} + W_t - W_t^{\text{match}}}.$$

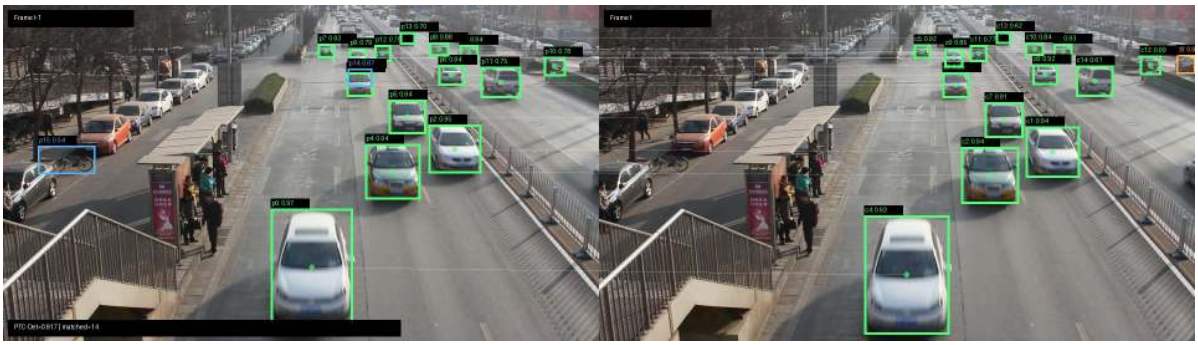


Figure 3.2.: Visualization example of the proposed PTC-Det consistency measurement between two adjacent video frames. Green bounding boxes indicate detections that are successfully matched across the two frames through class-aware IoU matching, while blue and orange bounding boxes represent unmatched detections that only appear in frame $t-1$ and frame t, respectively. The connecting lines indicate matched detection correspondences, where the line labels denote IoU scores and the bounding-box labels denote confidence scores. For each matched pair (i, j), the confidence contribution is computed as $w_{ij} = \min(s_i^{t-1}, s_j^t)$ and the final score is calculated using the confidence-weighted Jaccard similarity. In this example, most objects remain consistently detected across adjacent frames, resulting in a relatively high PTC-Det score.

This formalization extends the count-based detection consistency score by considering the confidence of each detection. Compared to the simply counted detection results, this confidence-weighted formulation reduces the influence of low-confidence unstable predictions and focuses more on reliable detection results.

Higher PTC-Det_t values indicate that the detections are consistently produced in adjacent frames, while lower scores suggest that detections become unstable and may imply a higher frame-level difficulty.

3.3.2. Association Consistency (PTC-Ass)

Association consistency measures whether the detections can be clearly and unambiguously associated between adjacent video frames. In difficult frames, due to factors such as crowding, occlusion or object interactions, multiple detections may produce similar matching candidates, thereby making the temporal correspondence relationship ambiguous. Therefore, unstable association relationships can reflect higher frame-level difficulty.

For each detection d_j^t in the current frame, the best matched detection in the previous frame is determined based on the Intersection-over-Union (IoU) value:

$$s_1(j) = \max_i \text{IoU}(d_i^{t-1}, d_j^t),$$

while the second-best matching score is defined as:

$$s_2(j) = \max_{i \neq m(j)} \text{IoU}(d_i^{t-1}, d_j^t),$$

where $m(j)$ denotes the index of the best matched detection.

Based on the difference between the best and second-best matching scores, the association consistency score for each matched detection is defined as:

$$a_j = \frac{s_1(j) - s_2(j)}{s_1(j) + \epsilon},$$

where ϵ is a small constant for numerical stability.

To further reduce the influence of unreliable detections, the frame-level association consistency score is computed using confidence-weighted averaging:

$$\text{PTC-Ass}_t = \frac{\sum_{j=1}^{M_t} s_j^t a_j}{\sum_{j=1}^{M_t} s_j^t},$$

where M_t denotes the number of matched detections between adjacent frames and s_j^t represents the confidence score of the current-frame detection.

This formulation measures the ambiguity of temporal association. If the best matched detection is significantly better than the second-best candidate, the association is considered stable and unambiguous. In contrast, when multiple detections produce similar matching scores, the association becomes uncertain, which may indicate higher frame-level difficulty.

Therefore, lower PTC-Ass$_t$ scores are usually associated with crowded scenes, severe occlusion or complex object interactions, while higher scores indicate clearer and more stable temporal association consistency.

3.3.3. Localization Consistency (PTC-Loc)

Localization consistency refers to whether object localization remains temporally stable across adjacent video frames. In difficult frames, due to factors such as motion blur, abrupt object movement, inaccurate regression or occlusion, localization may become unstable. Therefore, unstable localization behavior can be regarded as an indicator of higher frame-level difficulty.

Unlike directly measuring the Intersection-over-Union (IoU) between adjacent detections, the proposed localization consistency formulation additionally considers the temporal continuity of motion. For a matched detection pair (d_i^{t-1}, d_j^t), the object location in frame t is first predicted using the motion estimated from previous frames.

Let the center coordinates of a bounding box be denoted as:

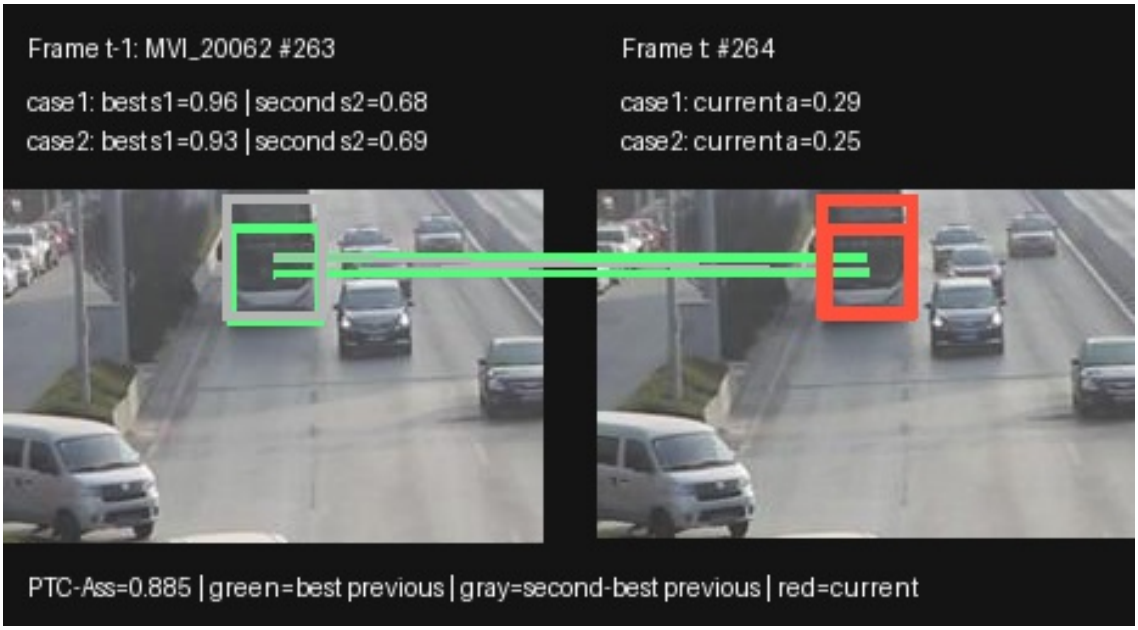


Figure 3.3.: Visualization of the proposed PTC-Ass score on two adjacent frames. The right crop shows the current frame t, while the left crop shows the previous frame $t-1$. For each current detection d_j^t, the best and second-best matching detections in the previous frame are computed by IoU: $s_1(j) = \max_i \mathrm{IoU}(d_i^{t-1}, d_j^t)$ and $s_2(j) = \max_{i \neq m(j)} \mathrm{IoU}(d_i^{t-1}, d_j^t)$, where $m(j)$ is the index of the best match. The association consistency of the current detection is then defined as $a_j = (s_1(j) - s_2(j))/(s_1(j) + \epsilon)$. In this example, the same bus is detected by two highly overlapping boxes in the previous frame. As a result, the current bus detection has both a strong best match and a non-negligible second-best match (e.g., $s_1 = 0.96$, $s_2 = 0.68$), which reduces a_j and indicates ambiguous temporal association. The frame-level PTC-Ass score is obtained by confidence-weighted averaging over matched current detections: $\text{PTC-Ass}_t = \frac{\sum_{j=1}^{M_t} s_j^t a_j}{\sum_{j=1}^{M_t} s_j^t}$. Green boxes/lines denote the best previous match, gray boxes/lines denote the second-best previous match and red boxes denote the current detections.

$$c = (x, y).$$

For a matched object trajectory across frame $t-2$ and frame $t-1$, the motion vector is estimated as:

$$\Delta c_i^{t-1} = c_i^{t-1} - c_i^{t-2}.$$

The predicted center position in frame t is then computed as:

$$\hat{c}_i^t = c_i^{t-1} + \Delta c_i^{t-1}.$$

Using the predicted center position together with the estimated bounding-box scale change, a predicted bounding box $\hat{b}_i^t$ is constructed.

The localization overlap consistency between the predicted box and the current detection is measured by:

$$\mathrm{IoU}(\hat{b}_i^t, b_j^t).$$

To further evaluate spatial motion smoothness, a DIoU-like center similarity between the predicted position and the observed position is defined as:

$$\mathrm{CS}(i,j) = \exp\left(-\frac{(\hat{c}_{i,x}^t - c_{j,x}^t)^2 + (\hat{c}_{i,y}^t - c_{j,y}^t)^2}{C^2(\hat{b}_i^t, b_j^t) + \epsilon}\right),$$

where $(\hat{c}_{i,x}^t, \hat{c}_{i,y}^t)$ denotes the center of the predicted box $\hat{b}_i^t$, $(c_{j,x}^t, c_{j,y}^t)$ denotes the center of the current detection box b_j^t and $C^2(\hat{b}_i^t, b_j^t)$ denotes the squared diagonal length of the smallest enclosing box covering these two boxes.

The localization consistency score for each matched detection pair is then computed by combining overlap consistency and center similarity:

$$l_j = \left(\mathrm{IoU}(\hat{b}_i^t, b_j^t)\right)^{\alpha} \cdot \left(\mathrm{CS}(i,j)\right)^{1-\alpha},$$

where α controls the balance between overlap consistency and motion smoothness.

Finally, the frame-level localization consistency score is computed using confidence-weighted averaging:

$$\text{PTC-Loc}_t = \frac{\sum_{j=1}^{M_t} s_j^t \, l_j}{\sum_{j=1}^{M_t} s_j^t},$$

where M_t denotes the number of matched detections between adjacent frames.

Compared with detection consistency and association consistency, localization consistency focuses more on the spatial stability of bounding box predictions and temporal coherence. A higher PTC-Loc$_t$ score indicates more stable localization behavior across adjacent frames, while a lower score may suggest unstable localization caused by complex scene conditions.

3.3.4. Composite Difficulty Score

The proposed PTC-IoU method decomposes frame-level temporal consistency into three interpretable sub-dimensions: detection consistency, association consistency and localization consistency. Since these sub-dimensions capture different aspects of temporal stability, a unified frame-level consistency score is constructed and used as a difficulty indicator.

Given the three temporal consistency scores:

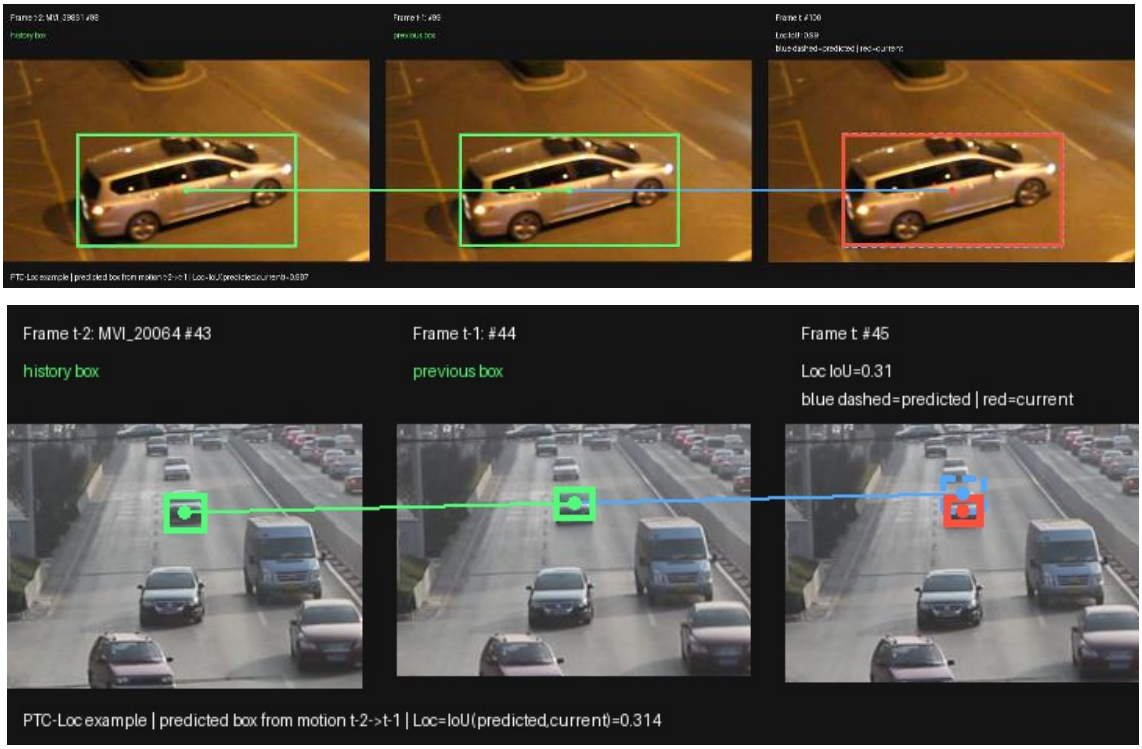


Figure 3.4.: Visualization of the proposed PTC-Loc score. The three crops correspond to frames $t-2$, $t-1$ and t. The green boxes show the matched detections in the two previous frames. Based on the displacement from frame $t-2$ to frame $t-1$, the expected location in frame t is predicted by motion extrapolation. The blue dashed box denotes the predicted location $\hat{d}^t$, while the red box denotes the actual current detection d^t. The localization consistency is computed using a DIoU-like formulation: PTC-Loc $= \mathrm{IoU}(\hat{d}^t, d^t)^\alpha \cdot \mathrm{CS}(\hat{d}^t, d^t)^{1-\alpha}$, where CS measures center consistency between the predicted and current detections. The upper example shows stable localization, where the predicted box and the current detection highly overlap and remain spatially consistent, resulting in a high score. The lower example shows unstable localization, where the current detection deviates from the predicted trajectory, leading to lower overlap and center consistency and thus a much lower score.

$$\text{PTC-Det}_t, \qquad \text{PTC-Ass}_t, \qquad \text{PTC-Loc}_t,$$

the final composite PTC-IoU score is computed using weighted aggregation:

$$\text{PTC-IoU}_t = \mathcal{F}\left(\text{PTC-Det}_t, \text{PTC-Ass}_t, \text{PTC-Loc}_t\right),$$

where $\mathcal{F}(\cdot)$ denotes the aggregation function.
In this work, multiple aggregation strategies are explored, including weighted arithmetic

mean, weighted geometric mean and weighted power mean. For the weighted arithmetic formulation, the composite score is defined as:

$$\text{PTC-IoU}_t = w_d \cdot \text{PTC-Det}_t + w_a \cdot \text{PTC-Ass}_t + w_l \cdot \text{PTC-Loc}_t,$$

subject to:

$$w_d + w_a + w_l = 1,$$

where w_d, w_a and w_l denote the weights of the three sub-dimensions.
For the weighted geometric formulation, the score is computed as:

$$\text{PTC-IoU}_t = \left(\text{PTC-Det}_t^{w_d} \cdot \text{PTC-Ass}_t^{w_a} \cdot \text{PTC-Loc}_t^{w_l}\right).$$

In addition, component-wise exponent transformation is further introduced:

$$\tilde{D}_t = \text{PTC-Det}_t^{\alpha}, \qquad \tilde{A}_t = \text{PTC-Ass}_t^{\beta}, \qquad \tilde{L}_t = \text{PTC-Loc}_t^{\gamma},$$

where α, β and γ control the contribution sensitivity of different sub-dimensions.

The final aggregation configuration is selected by grid search on the validation split. Algorithm 1 summarizes this procedure. For each candidate configuration, the three consistency sub-dimensions are first transformed by component-wise exponents and then combined using one of the candidate aggregation families, including weighted arithmetic mean, weighted geometric mean, weighted power mean and a gated formulation. The resulting frame-level PTC-IoU scores are compared with validation targets using ranking-oriented metrics and a combined validation score is computed.

After all candidate configurations have been evaluated, the configuration with the best combined score is finally identified. To avoid selecting an overly degenerate configuration when several candidates achieve nearly identical validation performance, a near-best selection rule is applied. Among configurations whose combined score is within a small tolerance of the raw best score, the final configuration is selected by preferring a more balanced use of the detection, association and localization sub-dimensions. This selection strategy does not force all sub-dimensions to contribute, but allows more balanced configurations when they introduce almost no validation-score loss.

Therefore, the aggregation strategy mainly serves as a practical mechanism for combining complementary temporal consistency information from different sub-dimensions.

3.4. Ranking-oriented Evaluation Perspective

The main objective of this method is not to precisely regress frame-level detection metrics, but to estimate the relative difficulty of video frames. In practical adaptive inference scenarios, the routing system mainly needs to establish a reliable ranking of difficulty between frames, in order to determine whether a particular frame should be handled by a weaker detector or a stronger one.

Algorithm 1 Grid search for PTC-IoU aggregation configuration

Require: Frame-level scores PTC-Det$_t$, PTC-Ass$_t$, PTC-Loc$_t$
Require: Validation targets $y_t \in \{\text{AP}, \text{Precision}, \text{Recall}\}$
Require: Candidate aggregation families $\mathcal{A}$ and parameter sets Φ_F
Require: Candidate exponent values $\mathcal{E}$
Ensure: Best aggregation configuration θ^*
Initialize result set $\mathcal{R} \leftarrow \emptyset$
for all aggregation family $F \in \mathcal{A}$ **do**
  for all exponent configuration $(\alpha, \beta, \gamma) \in \mathcal{E}$ **do**
    $\tilde{D}_t \leftarrow \text{PTC-Det}_t^{\alpha}$
    $\tilde{A}_t \leftarrow \text{PTC-Ass}_t^{\beta}$
    $\tilde{L}_t \leftarrow \text{PTC-Loc}_t^{\gamma}$
    for all candidate parameter configuration $\phi \in \Phi_F$ **do**
      $S_t \leftarrow F(\tilde{D}_t, \tilde{A}_t, \tilde{L}_t; \phi)$
      for all validation target y_t **do**
        Compute ranking metrics between $\{S_t\}$ and $\{y_t\}$
        Compute combined validation score C
        Add $(F, \alpha, \beta, \gamma, \phi, C)$ to $\mathcal{R}$
      end for
    end for
  end for
end for
$C^* \leftarrow \max_{\theta \in \mathcal{R}} C(\theta)$
$\mathcal{R}' \leftarrow \{\theta \in \mathcal{R} \mid C(\theta) \geq C^* - \epsilon\}$
$\theta^* \leftarrow$ candidate in $\mathcal{R}'$ with the best component-balance score
return θ^*

However, direct regression of frame-level metrics is itself a highly challenging task. In video object detection, frame-level metrics (such as AP, precision and recall) are often noisy and unstable due to the limited number of objects in a single frame. Moreover, frame-level AP values typically contain a large number of tied values, such as 0 or 1, which further reduces the reliability of direct numerical regression.

Therefore, this study mainly focuses on ranking-based evaluation rather than absolute metric prediction. The proposed method is assessed based on whether the generated temporal consistency scores can correctly distinguish relatively simple frames from difficult frames.

To evaluate the ranking ability of the proposed indicator, several ranking-oriented metrics were adopted.

Kendall Tau is used to measure the consistency between the predicted frame difficulty ordering and the target metric ordering:

$$\tau = \frac{N_c - N_d}{\frac{1}{2}n(n-1)},$$

where N_c and N_d denote the numbers of concordant and discordant frame pairs, respectively.

Spearman correlation is additionally used to evaluate monotonic ranking consistency:

$$\rho = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)},$$

where d_i denotes the rank difference between two ranking lists.

Furthermore, pairwise ranking accuracy is introduced to directly evaluate whether the proposed indicator can correctly determine the relative difficulty relationship between two frames. Given two frames (i, j), a prediction is considered correct if:

$$(\hat{s}_i - \hat{s}_j)(y_i - y_j) > 0,$$

where $\hat{s}$ denotes the predicted temporal consistency score and y denotes the target frame-level metric.

The pairwise accuracy is then computed as:

$$\text{PairwiseAcc} = \frac{N_{\text{correct}}}{N_{\text{valid}}},$$

where N_{correct} denotes the number of correctly ordered frame pairs and N_{valid} denotes the total number of evaluated frame pairs.

Compared with direct numerical regression methods, ranking-based evaluation is more aligned with the practical objectives of adaptive routing systems. Therefore, the proposed method mainly emphasizes the relative ranking ability for difficulty assessment based on temporal consistency, rather than precise frame-level metric prediction.

3.5. Routing Framework

The proposed temporal consistency framework is further applied to adaptive routing scenarios in video object detection. The main objective of this routing framework is to dynamically allocate different frames to detectors with varying computational costs based on the estimated frame difficulty.

In practical video analysis systems, not all frames require the same detection capability. Relatively simple frames can usually be adequately processed by lightweight detectors, while complex frames may require more powerful detectors with higher computational complexity. Therefore, adaptive routing based on frame-level difficulty assessment provides a potential approach for improving the balance between efficiency and performance in video object detection systems.

For a video sequence, frame-level detection results are first generated by a weaker detector. Then, the proposed PTC indicator is applied to the detection results of adjacent frames to calculate the corresponding temporal consistency score:

$$\text{PTC-IoU}_t.$$

Based on the estimated difficulty score, each frame is subsequently assigned to either the weak detector or the strong detector according to a predefined routing threshold.

Let θ denote the routing threshold. The routing strategy is defined as:

$$R_t = \begin{cases} \text{Weak Detector}, & \text{PTC-IoU}_t \geq \theta, \\ \text{Strong Detector}, & \text{PTC-IoU}_t < \theta. \end{cases}$$

Since lower temporal consistency scores usually indicate that frames are more difficult to process, frames with lower PTC-IoU values will be assigned to stronger detectors for more reliable predictions.

After the routing process, the final detection results are obtained by combining the output results of the two detectors based on the routing decisions.

To evaluate the effectiveness of the proposed routing framework, further studies were conducted under different routing budgets. In the routing-budget experiments, the threshold is selected such that a fixed proportion of frames is routed to the strong detector. The routing budget controls the proportion of frames processed by the strong detector. With a smaller budget, only a limited number of difficult frames can be reallocated, while a larger budget enables more frames to benefit from stronger detection capability.

The proposed routing framework was evaluated from both performance and efficiency perspectives. Detection accuracy is measured using frame-level and dataset-level evaluation metrics, while computational efficiency is analyzed based on the proportion of frames processed by the strong detector.

Compared with static detector selection methods, the proposed routing framework aims to dynamically allocate computational resources by estimating frame difficulty based on temporal consistency, thereby improving the balance between the overall efficiency and performance of video object detection systems.

4. Experimental Setup

4.1. Experimental Configuration

The overall experimental flow adopted in this thesis is shown in Figure 4.2. All experiments are carried out on the MOT17 dataset [23], using a video-level train-test split.

Given an input video sequence, a weak detector is first applied to each frame to generate object detection results, including object category, bounding box position and confidence score. Based on the detection outputs of adjacent frames, the proposed PTC-IoU method calculates the frame-level difficulty score by analyzing detection consistency, association consistency and localization consistency over time.

As a comparison, representative quality estimation baselines including IoU-Net [12] and GFL [13] are also applied to generate frame-level quality scores. IoU-Net evaluates localization quality through the predicted IoU value, while GFL uses the Generalized Focal Loss representation to model detection quality. These baseline scores are then compared with the proposed PTC-IoU indicator in ranking-oriented and routing-oriented evaluations.

In the ranking-oriented setting, the estimated difficulty score is compared with frame-level detection performance. In the routing-oriented setting, the generated score is used to determine whether a frame should be processed by the weak detector or re-evaluated by a stronger detector. Subsequently, the generated detection outputs are used for further evaluation.

In addition, the computational overhead introduced by the proposed method is analyzed from the perspective of runtime. By comparing inference latency, throughput and relative computational cost, the practical applicability of the proposed method and the baseline methods in adaptive video object detection systems is evaluated.

Detailed experimental configurations, including dataset split, detector settings, baseline methods, evaluation metrics and hardware specifications, are summarized in Table 4.1.

The experiment aims to answer the following research questions:

RQ1: Can temporal consistency provide a meaningful frame-level difficulty estimate for video object detection?

RQ2: Can the estimated difficulty score improve adaptive detector routing under different computational budgets?

RQ3: How does the proposed PTC-IoU indicator compare with existing quality estimation methods in terms of computational efficiency?

To provide an intuitive impression of the video data used in the experiments, several representative frames from different MOT17 sequences are shown in Figure 4.1. These examples illustrate typical pedestrian detection scenarios, including crowded scenes, camera-view changes and different object scales.

Table 4.1.: Experimental configuration used throughout this thesis.

Item	Value
Dataset	MOT17 (pedestrian).
MOT17 Split 1	Seed 42, by sequence, approximately 60:40 frames. Training sequences: MOT17-04-FRCNN, MOT17-05-FRCNN, MOT17-09-FRCNN, MOT17-10-FRCNN. Validation sequences: MOT17-02-FRCNN, MOT17-11-FRCNN, MOT17-13-FRCNN.
MOT17 Split 2	Seed 24, by sequence, approximately 60:40 frames. Training sequences: MOT17-02-FRCNN, MOT17-04-FRCNN, MOT17-05-FRCNN, MOT17-13-FRCNN. Validation sequences: MOT17-09-FRCNN, MOT17-10-FRCNN, MOT17-11-FRCNN.
Weak Detector	Faster R-CNN MobileNet V3 Large 320 FPN [2].
Strong Detector	Faster R-CNN ResNet50 FPN V2 [2].
Method	PTC-IoU (ours).
Baselines	IoU-Net, GFL.
Routing References	Random Routing, Oracle Routing.
Ranking Metrics	Pairwise Accuracy, Kendall Tau, Pearson Correlation, Spearman Correlation.
Routing Metrics	AP, Precision, Recall.
Hardware	AMD EPYC 7302 16-Core Processor, NVIDIA A40 GPU (46 GB VRAM), 251 GB RAM.

(a) MOT17-02-FRCNN (b) MOT17-09-FRCNN (c) MOT17-11-FRCNN

Figure 4.1.: Example frames from three different MOT17 video sequences. The selected frames show different pedestrian detection conditions and provide visual examples of the dataset used in the experiments.

4.2. Evaluation Objectives

To better understand the role of frame-level difficulty estimation, we evaluated the generated scores from three complementary perspectives: ranking quality, routing effectiveness and computational performance. As shown in Figure 4.3, the estimator assigned a difficulty score to each video frame. Subsequently, the quality of the score was analyzed based on whether it preserved the relative ordering of frame difficulty, whether it supported effective adaptive routing and whether it introduced only limited computational overhead.

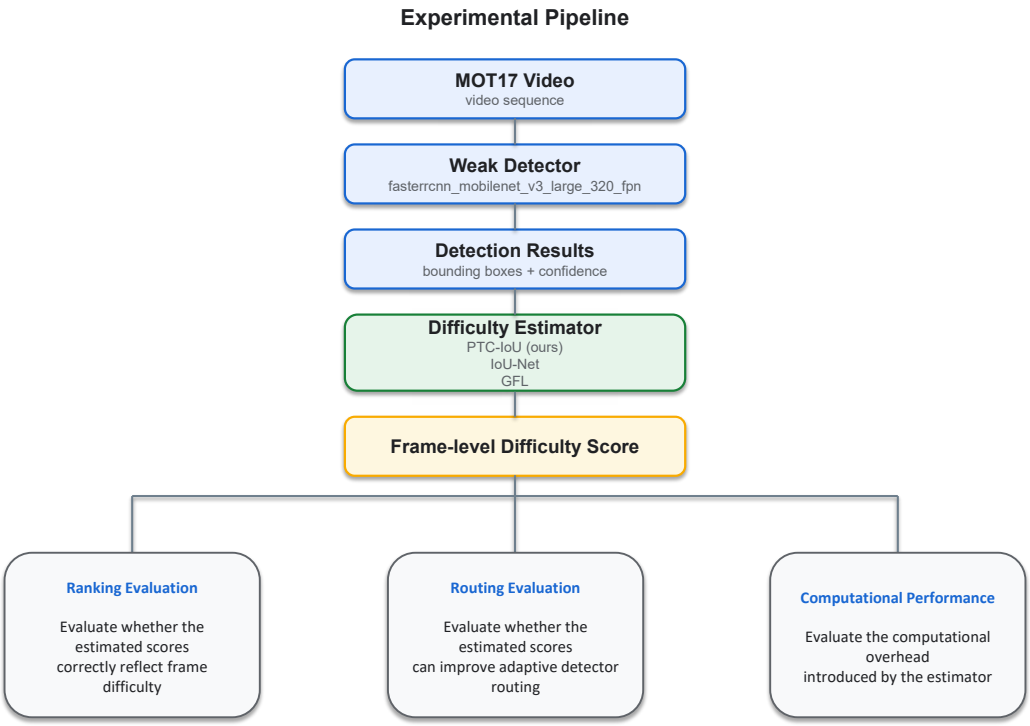


Figure 4.2.: The overall experimental process of the proposed method. The frame-level difficulty score is generated from the weak detector outputs and then evaluated from three perspectives: ranking, routing and computational performance.

4.2.1. Ranking-oriented Evaluation

The primary goal of this thesis is to study whether temporal consistency can provide meaningful signals for frame-level difficulty evaluation. Since adaptive routing mainly relies on the relative order between frames rather than precise metric prediction, the proposed indicator is mainly evaluated from the ranking perspective.

For each frame, the generated temporal consistency score is compared with the corresponding frame-level detection performance. The consistency between the estimated difficulty scores and the performance-based difficulty values is measured using pairwise accuracy, Kendall Tau, Pearson correlation and Spearman correlation. Higher values indicate a stronger ability to distinguish between relatively easy and difficult frames.

4.2.2. Routing-oriented Evaluation

The second goal is to evaluate whether temporal consistency can be effectively applied to adaptive detector routing. In this scenario, the frame-level difficulty score generated from the weak detector outputs is used to determine which frames should be further processed by the

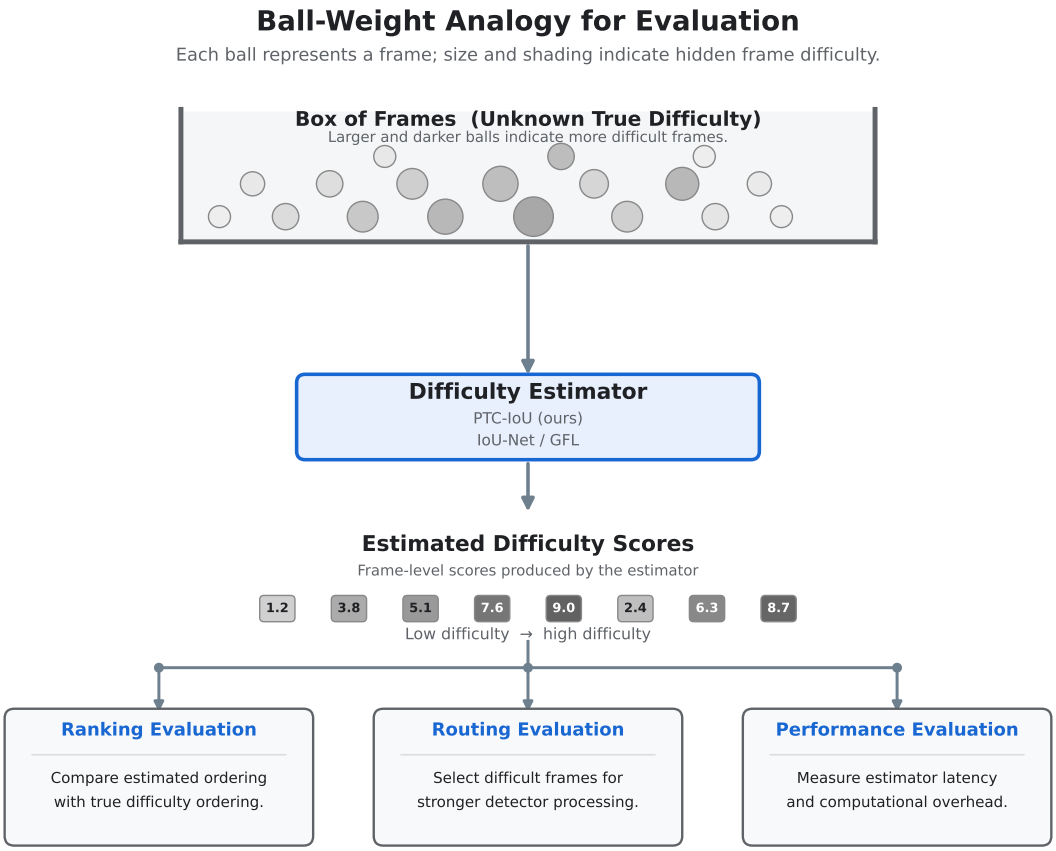


Figure 4.3.: Overview of the evaluation objectives for frame-level difficulty estimation. The estimator assigns a difficulty score to each video frame and the score is then evaluated from three aspects: ranking quality, routing effectiveness and computational performance.

strong detector.

In order to simulate different computational constraints, a routing budget ρ is introduced, indicating the proportion of frames re-evaluated by the strong detector. When $\rho = 0$, all frames are processed by the weak detector; when $\rho = 1$, all frames are processed by the strong detector. A ρ value between these two extremes represents a mixed setting, where only part of the frames are selected for strong-detector inference.

The generated difficulty score is used to rank the frames according to the estimated difficulty. The routing strategy is based on the assumption that frames with lower weak-detector quality are more likely to achieve larger performance improvements when re-evaluated by the strong detector. Therefore, frames estimated as more difficult are considered to have higher potential gain and are prioritized for strong-detector processing.

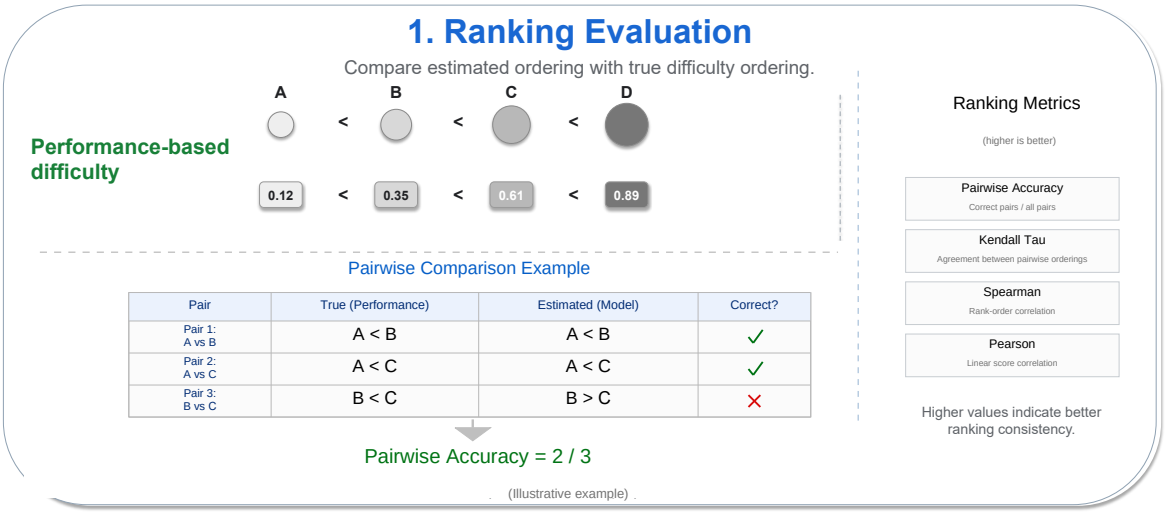


Figure 4.4.: Schematic diagram of ranking-oriented evaluation. The estimated frame difficulty score is compared with the performance-based difficulty ranking. High ranking consistency indicates that the estimator reflects the relative difficulty between frames more accurately.

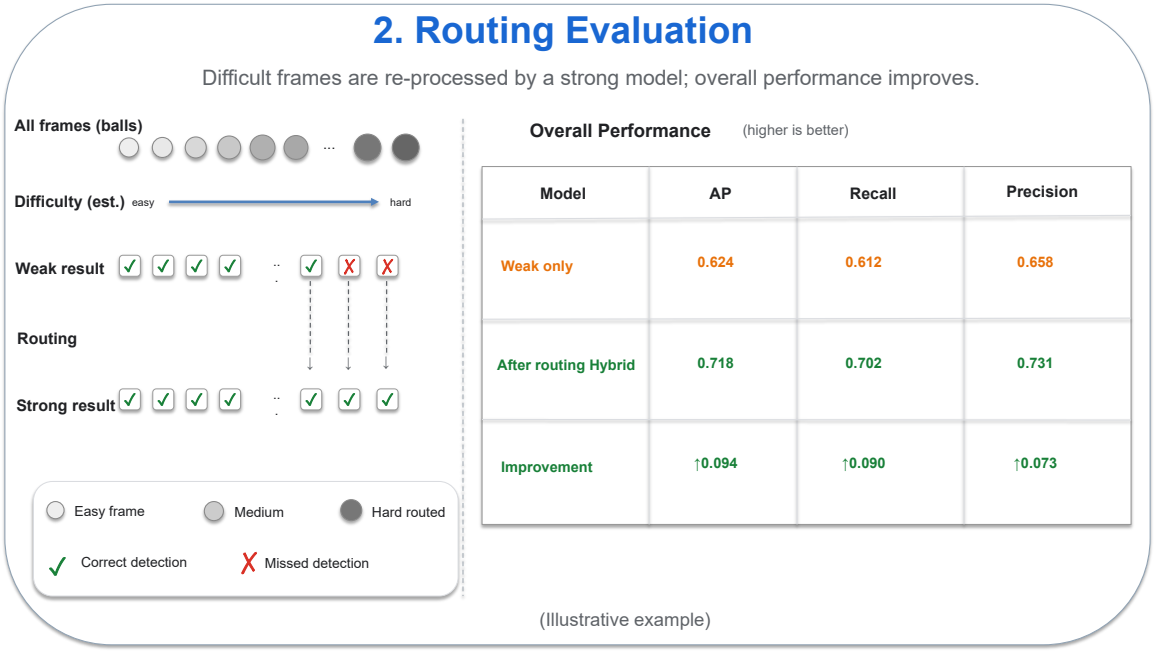


Figure 4.5.: Schematic diagram of routing-oriented evaluation. Frames that are estimated to be difficult are selectively processed by a stronger detector. The final hybrid detection results are evaluated by Average Precision (AP), precision and recall to measure the effectiveness of the routing strategy.

Frames with the lowest estimated consistency, namely the highest estimated difficulty, are routed to the strong detector until the specified routing budget is reached. The final detection performance is evaluated by Average Precision (AP), precision and recall [1].

As a comparison, two reference routing strategies are also considered. Random routing uniformly selects frames for strong-detector processing and provides a lower reference line. Oracle routing assumes that the true frame-level detection difficulty can be obtained, so it represents the performance upper bound achievable under a given routing budget.

A successful routing strategy should allocate limited computational resources to the most difficult frames and achieve higher detection performance than the baseline methods.

4.2.3. Computational Performance

In addition to evaluating ranking quality and routing effectiveness, the computational overhead of the proposed method is also considered. Since temporal consistency is designed as a post-model difficulty estimation method, an important goal is to provide meaningful difficulty assessment while introducing minimal additional computation.

Unlike IoU-Net and GFL, which require additional learned components and inference operations, the proposed PTC-IoU indicator estimates frame difficulty directly from temporal consistency cues without any additional training. Therefore, a key goal is to maximize the performance improvement brought by adaptive routing while reducing computational overhead as much as possible.

The proposed method is compared with IoU-Net and GFL in terms of runtime efficiency. As summarized in Table 4.3, computational performance is analyzed in terms of latency, throughput, runtime overhead and relative slowdown. These criteria are used to evaluate the trade-off between routing effectiveness and computational cost. A practical difficulty estimator should achieve meaningful performance improvements while introducing limited additional runtime overhead, thus providing an efficient solution for adaptive video object detection systems.

Table 4.2.: Runtime measurement protocol for the computational performance evaluation.

Item	Value
Evaluation setting	Online frame-by-frame inference from preloaded image tensors to scores.
Input processing	Images are resized to 640×384 and preloaded before timing; disk I/O and image loading are excluded.
Dataset	MOT17 sequences MOT17-02-FRCNN, MOT17-05-FRCNN, MOT17-10-FRCNN and MOT17-13-FRCNN.
Frames	First 500 frames per sequence.
Batch size	1.
Compared methods	Weak detector, strong detector, PTC-IoU, GFL and IoU-Net.
Baseline timing	GFL and IoU-Net include the weak detector forward pass, additional RoI feature extraction and head inference.
Repeated runs	7 runs per method; the first 2 runs are discarded as warm-up.
Hardware	NVIDIA A40 GPU.

Table 4.3.: Evaluation criteria used for computational performance analysis.

Criterion	Description	Preferred Direction
Latency	Average processing time per frame	Lower is better
Throughput (FPS)	Number of frames processed per second	Higher is better
Runtime overhead	Additional latency compared with the weak detector	Lower is better
Slowdown	Relative runtime compared with the weak detector	Lower is better

5. Evaluation

5.1. Ranking Evaluation

This section evaluates whether the proposed PTC-IoU indicator can provide a meaningful frame-level difficulty estimate for video object detection. By using ranking-based evaluation metrics, including pairwise accuracy, Kendall Tau, Pearson correlation and Spearman correlation, the generated difficulty score is compared with frame-level detection performance. In addition, the relationship between the estimated score and detection quality is also analyzed to explore whether lower temporal consistency corresponds to lower detection quality.

5.1.1. Overall Ranking Performance

Table 5.1 compares the ranking performance of PTC-IoU on two validation splits with the learned quality estimation baselines GFL [13] and IoU-Net [12]. The results show that there are clear differences in relative ranking performance under different dataset splits and evaluation targets.

On Split 1, the learned baseline methods usually show stronger ranking performance. For Average Precision (AP), IoU-Net achieves the highest result across all ranking metrics, with a pairwise accuracy of 0.757, a Kendall Tau coefficient of 0.509, a Spearman correlation coefficient of 0.708 and a Pearson correlation coefficient of 0.697. Recall also shows a similar trend, where IoU-Net again achieves the best performance. For precision, GFL performs slightly better than IoU-Net and achieves the strongest ranking result on this split. In contrast, on Split 1, the correlation values of PTC-IoU are lower, especially in terms of Pearson correlation. This indicates that under this split, the learned quality estimation methods can provide more direct ranking signals for frame-level AP, precision and recall.

However, the results on Split 2 show a different trend. On Split 2, PTC-IoU achieves the best ranking performance for three evaluation targets on most ranking metrics. For Average Precision (AP), PTC-IoU obtains a pairwise accuracy of 0.768, a Kendall Tau coefficient of 0.531, a Spearman correlation coefficient of 0.738 and a Pearson correlation coefficient of 0.694, outperforming both GFL and IoU-Net. Precision and recall also show similar improvements, where PTC-IoU reaches the highest values in pairwise accuracy, Kendall Tau and Spearman correlation. For precision, IoU-Net is only slightly better than PTC-IoU in terms of Pearson correlation, with values of 0.680 and 0.679, respectively.

These results show that PTC-IoU is not better than the learned quality estimation baselines in all scenarios. Its effectiveness depends on the dataset split and the distribution of the underlying frames. However, the strong performance on Split 2 shows that temporal consistency can provide a meaningful ranking signal for frame-level difficulty estimation.

Table 5.1.: Ranking performance comparison between the proposed PTC-IoU indicator and representative quality estimation baselines on two validation splits. Higher values indicate better agreement with the corresponding frame-level performance ordering.

Target	Split	Method	Pairwise ↑	Kendall τ ↑	Spearman ρ ↑	Pearson r ↑
AP	Split 1	GFL	0.748	0.487	0.679	0.667
		IoU-Net	**0.757**	**0.509**	**0.708**	**0.697**
		PTC-IoU (Ours)	0.709	0.406	0.568	0.396
	Split 2	GFL	0.657	0.315	0.464	0.451
		IoU-Net	0.706	0.397	0.581	0.554
		PTC-IoU (Ours)	**0.768**	**0.531**	**0.738**	**0.694**
Precision	Split 1	GFL	**0.827**	**0.643**	**0.837**	**0.821**
		IoU-Net	0.813	0.621	0.817	0.810
		PTC-IoU (Ours)	0.772	0.526	0.696	0.468
	Split 2	GFL	0.734	0.458	0.638	0.669
		IoU-Net	0.746	0.485	0.677	**0.680**
		PTC-IoU (Ours)	**0.771**	**0.524**	**0.719**	0.679
Recall	Split 1	GFL	0.736	0.464	0.655	0.660
		IoU-Net	**0.740**	**0.475**	**0.668**	**0.671**
		PTC-IoU (Ours)	0.682	0.354	0.508	0.353
	Split 2	GFL	0.662	0.315	0.457	0.449
		IoU-Net	0.709	0.399	0.576	0.553
		PTC-IoU (Ours)	**0.771**	**0.535**	**0.736**	**0.708**

Unlike IoU-Net and GFL, PTC-IoU does not require additional learned quality prediction heads. Therefore, even if its ranking performance is not always the highest, it still provides a lightweight and training-free alternative for estimating frame-level difficulty from temporal detection behavior.

5.1.2. Component Analysis

Table 5.2 further analyzes the contributions of each PTC component: detection consistency, association consistency and localization consistency. This analysis aims to understand which temporal consistency sub-dimensions have the strongest effect on improving frame-level ranking performance.

In Split 1, PTC-Ass is the dominant individual component. For AP, PTC-Ass achieves the best Kendall Tau, Spearman correlation and Pearson correlation among all components, while the pairwise accuracy of PTC-IoU is slightly higher. A similar situation occurs for precision and recall. Specifically, for precision, PTC-Ass has the highest Spearman correlation and Pearson correlation, while the complete PTC-IoU score slightly improves pairwise accuracy and Kendall Tau. For recall, PTC-Ass is almost the same as PTC-IoU. PTC-Ass is slightly

stronger in pairwise accuracy, Kendall Tau and Spearman correlation, while PTC-IoU is slightly stronger in Pearson correlation.

This shows that association consistency provides the most informative ranking signal on the first split. For the MOT17 dataset, this observation is reasonable, because pedestrian scenes usually contain challenges such as crowding, occlusion and identity ambiguity. In this case, the instability of associations between adjacent frames may strongly reflect frame difficulty. When detected objects become difficult to associate over time, the corresponding frame is more likely to contain challenging visual conditions.

In Split 2, the full PTC-IoU score is more stable and favorable. For AP, PTC-IoU performs best in terms of pairwise accuracy, Kendall Tau and Spearman correlation, while PTC-Loc achieves the highest value in Pearson correlation. For precision and recall, PTC-IoU shows the best performance in almost all ranking metrics. This shows that combining detection, association and localization consistency can provide a more robust difficulty signal than relying on a single temporal sub-dimension.

The results also show that different components can capture different aspects of frame-level difficulty. When association ambiguity is the main source of difficulty, PTC-Ass performs particularly well. In some cases, PTC-Loc can achieve strong Pearson correlation, indicating that localization consistency may better reflect continuous changes in frame-level quality. PTC-Det also performs well on Split 2, indicating that detection consistency has reference value when confidence stability and detection existence are closely related to frame-level performance. Overall, the component analysis supports the design of PTC-IoU as a combination of multiple complementary temporal consistency cues, rather than a single isolated signal.

5.1.3. Discussion

Ranking evaluation is included because adaptive routing relies on the relative ordering generated by the difficulty estimator. Under a given computational budget, the strong detector can only process a subset of frames and these frames are usually selected based on their estimated difficulty scores. Therefore, before analyzing the final routing performance, it is necessary to verify whether the estimator can assign lower scores to relatively more difficult frames. If the score cannot preserve the relative difficulty order between frames, it is unlikely to provide a reliable basis for budget-based routing or offloading decisions.

The ranking evaluation indicates that the temporal consistency method can provide meaningful frame-level difficulty estimates, but its performance differs from learning-based quality estimation baselines. IoU-Net and GFL demonstrate strong ranking performance on Split 1, while PTC-IoU achieves the best results for most targets and ranking metrics on Split 2. This difference suggests that ranking quality is sensitive to the dataset split and the specific frame types included in the validation sequences.

An important observation is that high ranking performance does not necessarily imply stronger routing effectiveness. Ranking metrics evaluate whether the ordering generated by the estimator is consistent with frame-level AP, precision or recall. However, adaptive routing depends not only on whether difficult frames can be identified, but also on whether these frames can benefit from reprocessing by the strong detector. Some frames may remain

Table 5.2.: Component analysis of the proposed PTC-IoU indicator on two validation splits. Higher values indicate better ranking performance with respect to the corresponding frame-level evaluation target.

Target	Split	Component	Pairwise ↑	Kendall τ ↑	Spearman ρ ↑	Pearson r ↑
AP	Split 1	PTC-Det	0.631	0.260	0.375	0.190
		PTC-Ass	**0.709**	**0.407**	**0.570**	0.393
		PTC-Loc	0.629	0.248	0.360	0.213
		PTC-IoU (Ours)	**0.709**	0.406	0.568	**0.396**
	Split 2	PTC-Det	0.762	0.512	0.716	0.661
		PTC-Ass	0.729	0.450	0.641	0.620
		PTC-Loc	0.760	0.518	0.724	**0.703**
		PTC-IoU (Ours)	**0.768**	**0.531**	**0.738**	0.694
Precision	Split 1	PTC-Det	0.703	0.392	0.541	0.296
		PTC-Ass	0.771	0.525	**0.698**	**0.503**
		PTC-Loc	0.688	0.370	0.506	0.332
		PTC-IoU (Ours)	**0.772**	**0.526**	0.696	0.468
	Split 2	PTC-Det	0.753	0.494	0.686	0.629
		PTC-Ass	0.762	0.496	0.683	0.670
		PTC-Loc	0.731	0.464	0.655	0.625
		PTC-IoU (Ours)	**0.771**	**0.524**	**0.719**	**0.679**
Recall	Split 1	PTC-Det	0.608	0.212	0.307	0.144
		PTC-Ass	0.682	**0.355**	**0.510**	0.350
		PTC-Loc	0.605	0.202	0.294	0.167
		PTC-IoU (Ours)	**0.682**	0.354	0.508	**0.353**
	Split 2	PTC-Det	0.766	0.517	0.716	0.668
		PTC-Ass	0.728	0.442	0.626	0.612
		PTC-Loc	0.763	0.521	0.723	**0.709**
		PTC-IoU (Ours)	**0.771**	**0.535**	**0.736**	0.708

difficult for both the weak detector and the strong detector, while other frames may not appear to be the most difficult under weak-detector performance but can still be significantly improved by the strong detector.

Therefore, the ranking results should be understood as an intermediate evaluation of difficulty estimation quality, rather than a direct measurement of routing effectiveness. The fact that PTC-IoU is not always better than learned baseline methods in ranking performance does not necessarily mean that its routing performance will also be worse. Instead, the ranking evaluation shows that PTC-IoU can capture meaningful difficulty patterns from temporal detection behavior and the subsequent routing evaluation further explores whether these patterns can be transformed into actual performance improvements under different computational budgets.

Overall, the results give a positive answer to the first research question: temporal consistency

can provide useful frame-level difficulty signals for video object detection. At the same time, the split-dependent behavior and the difference between ranking and routing also show that an independent routing-oriented evaluation is necessary.

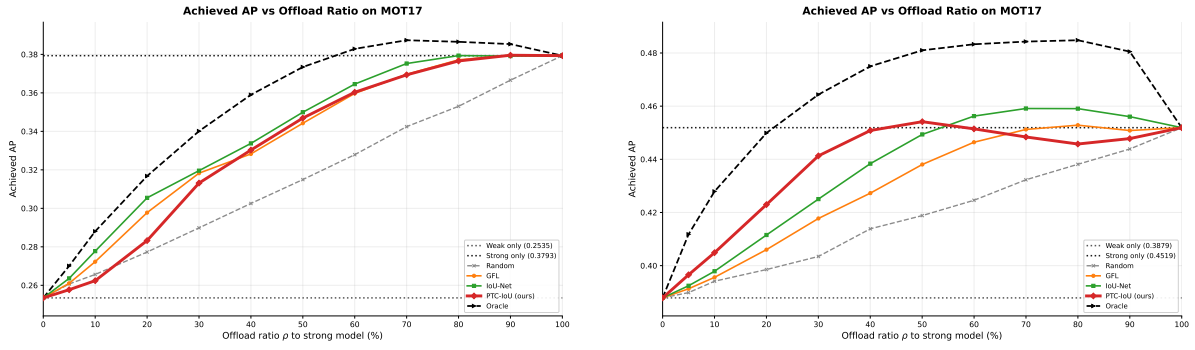
5.2. Routing Evaluation

This section evaluates whether the proposed PTC-IoU indicator can improve adaptive detector routing performance under different computational budgets. In this setting, the frame-level difficulty score is used to determine which frames should be handled by the strong detector. Frames that are estimated to be more difficult are re-evaluated by the strong detector first, while the remaining frames keep the output results of the weak detector. For PTC-IoU, lower temporal consistency indicates higher estimated difficulty; therefore, frames with lower PTC-IoU scores are prioritized for strong-detector processing.

The routing budget ρ represents the proportion of frames processed by the strong detector. When $\rho = 0$, all frames are processed by the weak detector; when $\rho = 1$, all frames are processed by the strong detector. Intermediate values correspond to mixed routing settings. Detection performance is evaluated by Average Precision (AP), precision and recall. The proposed method is compared with GFL, IoU-Net, random routing and oracle routing. Random routing provides a lower-bound reference under the same budget, while oracle routing is based on the true frame-level improvement and represents an upper-bound strategy.

5.2.1. AP-oriented Routing Performance

Figure 5.1 shows the performance of AP-oriented routing on two validation splits. The horizontal axis represents the offload ratio to the strong detector and the vertical axis represents the AP achieved after routing. The weak-only and strong-only lines represent the two endpoint settings, while the routing curves show how different estimators allocate the limited strong-detector budget.



(a) Split 1 (b) Split 2

Figure 5.1.: AP-oriented routing performance on two validation splits. The curves show the achieved AP under different offload ratios to the strong detector.

Table 5.3.: Routing performance summary at routing budget $\rho = 0.3$ on two validation splits. AP, precision, and recall are reported for the final routed detection outputs. Higher values are better.

Method	Split 1			Split 2		
	AP ↑	Precision ↑	Recall ↑	AP ↑	Precision ↑	Recall ↑
Weak only	0.253	0.165	0.309	0.388	0.210	0.447
Random routing	0.290	0.193	0.356	0.403	0.227	0.467
GFL	0.318	0.227	0.383	0.418	0.237	0.487
IoU-Net	**0.320**	**0.227**	**0.384**	0.425	0.240	0.497
PTC-IoU (Ours)	0.313	0.222	0.376	**0.441**	**0.252**	**0.514**
Oracle	0.340	0.232	0.402	0.464	0.257	0.529
Strong only	0.379	0.265	0.468	0.452	0.270	0.519

In the two split settings, as the routing budget increases, all non-random routing methods outperform the weak-only baseline, indicating that frame-level difficulty estimation can guide adaptive detector allocation. Compared with random routing, under the same offload ratio to the strong detector, the learned baselines and the proposed PTC-IoU indicator can usually achieve higher AP, especially under low-budget and medium-budget conditions. This shows that the selected frames are not chosen randomly, but contain more useful information for strong-detector processing.

There are differences in the performance of PTC-IoU on the two validation splits. On Split 1, PTC-IoU achieves a clear improvement in the low-to-medium budget range and remains competitive with GFL and IoU-Net. Around the medium routing ratio, the PTC-IoU curve approaches the strong-detector endpoint while only using the strong detector for part of the frames. On Split 2, the performance gap between the learned methods and the temporal-consistency-based routing method is relatively small, but PTC-IoU still follows the general trend of AP improvement as the strong-detector budget increases.

An important observation is that routing performance is not fully consistent with the ranking performance discussed before. Although learned quality estimators can achieve good ranking metrics, the routing curves show that effective detector allocation also depends on whether the selected frames actually benefit from strong-detector reprocessing. Therefore, AP-oriented routing provides a more direct evaluation method for measuring the practical utility of difficulty estimators in adaptive inference.

5.2.2. Summary across Detection Metrics

While Figure 5.1 focuses on AP, Table 5.3 summarizes routing performance across AP, Precision and Recall at a fixed routing budget. This provides a more compact comparison of the final routed detection outputs beyond a single metric.

The summary table shows that adaptive routing not only changes Average Precision (AP), but also affects the balance between precision and recall. Since routing replaces the selected

weak-detector predictions with the prediction results of the strong detector, the improvements in these metrics may be different. A method that improves AP may also improve recall by recovering missed detections or improve precision by replacing unreliable weak-detector outputs with more reliable strong-detector predictions.

Compared with random routing, estimation-based routing strategies usually achieve better detection performance under the same budget. This confirms that the estimated frame-level difficulty score is useful for selecting frames for strong-detector processing. PTC-IoU remains competitive with the learned baseline methods without requiring an additional quality prediction network. This is important because PTC-IoU is calculated directly from the weak detector outputs through temporal consistency cues, making it a lightweight post-model routing signal.

Oracle results provide an upper-bound reference for routing under the same budget. The gap between the proposed method and Oracle shows that there is still room for improvement in frame selection. However, the gap between random routing and the proposed method shows that temporal consistency already provides a meaningful signal for adaptive detector allocation.

5.2.3. Discussion

The routing evaluation answers the second research question of this thesis: can temporal consistency scores improve adaptive detector routing performance under a limited computational budget? The results show that, compared with weak-detector inference only and random routing, PTC-IoU can guide the strong detector to process selected frames and improve detection performance.

The comparison also highlights the difference between ranking-based and routing-based evaluation. Ranking metrics measure the quality of the estimator in ordering frames according to frame-level performance. However, routing performance depends on the actual gain obtained after replacing weak-detector outputs with strong-detector outputs. Some frames may remain difficult for both detectors, so the routing benefit is limited; while other frames, even if they are not ranked as the most difficult ones, may still be significantly improved by using the strong detector. Therefore, ranking quality is related to routing effectiveness, but they are not equivalent.

Overall, the routing results show that PTC-IoU can be used as a lightweight routing signal for adaptive video object detection and has practical value. It does not require additional training and can be calculated from temporal consistency patterns in weak-detector predictions. The next section further evaluates whether this advantage can be achieved with low computational overhead.

5.3. Computational Performance

Runtime evaluation is carried out in an online frame-by-frame setting. Each frame is independently processed from preloaded image tensors until the difficulty score is generated. This

Table 5.4.: Computational performance comparison on MOT17. Runtime is measured in an online frame-by-frame setting from preloaded image tensors to scores. For GFL and IoU-Net, the runtime includes the additional RoI feature extraction step used by their original scripts.

Method	ms/frame ↓	FPS ↑	Overhead ↓	Slowdown ↓
Weak only	17.25 ± 0.53	58.05 ± 1.86	0.00 ± 0.00	1.00x
PTC-IoU (Ours)	18.29 ± 0.44	54.70 ± 1.33	1.04 ± 0.24	1.06x
GFL	25.65 ± 0.48	39.00 ± 0.75	8.40 ± 0.10	1.49x
IoU-Net	25.82 ± 0.46	38.75 ± 0.70	8.57 ± 0.10	1.50x
Strong only	42.04 ± 0.21	23.79 ± 0.12	24.79 ± 0.41	2.44x

setting reflects the practical application scenario of adaptive video object detection, where a routing decision must be made before deciding whether a frame should be processed by the strong detector. Therefore, the reported latency mainly focuses on the computational cost of score generation, rather than image loading or disk I/O overhead.

As shown in Table 5.4, the weak detector achieves the lowest latency among the detector-only baselines, with 17.25 ms per frame. The proposed PTC-IoU indicator slightly increases the latency to 18.29 ms per frame, corresponding to an additional overhead of 1.04 ms per frame. This shows that the additional cost introduced by temporal consistency computation is limited.

In contrast, IoU-Net and GFL require 25.82 ms and 25.65 ms per frame, respectively. Although both methods are based on the same weak detector, they introduce additional computation through RoI feature extraction and learned quality-head inference [2, 12, 13]. Therefore, their overhead relative to the weak detector is about 8.56 ms and 8.40 ms per frame, which is significantly higher than that of PTC-IoU. This shows that the proposed training-free temporal consistency estimator provides a lighter difficulty estimation mechanism compared with learned quality prediction heads.

The strong detector requires 42.04 ms per frame, which is about 2.44 times slower than the weak detector. This large latency gap motivates the use of adaptive routing: processing all frames with the strong detector would introduce high computational overhead, while selectively routing difficult frames can achieve a better balance between detection accuracy and runtime efficiency. Overall, the results show that PTC-IoU introduces the smallest score-estimation overhead among the compared difficulty estimation methods, while remaining much more efficient than full strong-detector inference.

6. Conclusion and Future Work

6.1. Conclusion

This thesis studies proxy-based temporal consistency as a lightweight method for estimating frame-level difficulty in video object detection. The proposed PTC-IoU method estimates frame difficulty through three complementary temporal sub-dimensions: detection consistency, association consistency and localization consistency. Based on these sub-dimensions, a composite difficulty score is constructed and used for ranking evaluation and adaptive detector routing.

The experimental results show that temporal consistency can provide useful information for identifying difficult video frames. In the ranking evaluation, the proposed method reveals meaningful relationships between temporal inconsistency and frame-level detection quality. In the routing evaluation, compared with random routing, PTC-IoU improves the allocation of strong-detector computation in the evaluated scenarios, indicating that temporal consistency can serve as an effective proxy signal for adaptive video object detection. Additionally, the computational performance analysis indicates that, in an online frame-by-frame processing scenario, PTC-IoU introduces only limited additional overhead compared with the weak detector. This suggests that the proposed method can provide a lightweight difficulty signal while retaining most of the efficiency advantage of the weak detector.

6.2. Limitations

Nevertheless, there are still several limitations:

- **Dataset sensitivity.** The effectiveness of PTC-IoU depends on the dataset. Since this method relies on temporal consistency between video frames, its performance may vary depending on the motion patterns, object density, camera perspective, occlusion level and detector behavior of different datasets. The consistency cues that are effective in pedestrian sequences such as MOT17 may not be equally effective in other scenarios, such as vehicle detection, aerial videos or sparse-object scenes.
- **Limited parameter and model generalization.** The experiments in this study were conducted under limited parameter settings and detector configurations. Parameters such as the detection confidence threshold, matching threshold and the selection of weak–strong detector pairs can affect both the estimated difficulty score and the final routing performance. In particular, the detection score threshold affects the number of retained candidate detections. A lower threshold can retain richer temporal information,

but it also increases the number of detection boxes used in the matching process, resulting in additional computational overhead. Conversely, a higher threshold can reduce runtime, but it may remove uncertain detections that are important for difficulty estimation. Therefore, the current results may not fully reflect the performance of PTC-IoU under different parameter settings, detector combinations and deployment environments.

- **Gap to oracle routing.** Although PTC-IoU improves upon random routing, its routing performance still falls short of the oracle upper bound. This indicates that the current difficulty score does not fully capture which frames can benefit most from reprocessing by the strong detector. Therefore, there is still room for improvement in enhancing the accuracy of difficulty estimation and the effectiveness of detector allocation.
- **Simple aggregation of consistency sub-dimensions.** The aggregation of detection, association and localization consistency is still relatively simple. The current method uses manually designed components and selects weights through grid search to combine these three sub-dimensions. Although this design is easy to interpret, it may limit the adaptability of PTC-IoU to different datasets, detector architectures and deployment scenarios.

6.3. Future Work

Given these limitations, future work could focus on the following areas:

- **Evaluation on more diverse datasets.** PTC-IoU should be evaluated and further optimized on a broader range of datasets and scenarios to better understand in which situations temporal consistency is beneficial and in which situations its reliability may be lower.
- **Systematic study of parameters and detector configurations.** More systematic experiments should be conducted on key experimental settings, such as confidence thresholds, matching thresholds, smoothing windows and different combinations of weak and strong detectors. This would help improve the robustness and generalization ability of the method under various computational constraints and deployment scenarios.
- **Improvement of PTC-IoU, its sub-dimensions and aggregation strategies.** The design of PTC-IoU and its detection, association and localization consistency sub-dimensions can be further optimized. Besides improving each sub-dimension itself, future work could also explore more adaptive aggregation strategies to dynamically adjust the contribution of each sub-dimension, rather than relying only on manually selected or grid-searched weights. Another promising direction is to study proxy signals that are more directly related to potential routing gain, rather than merely estimating frame-level difficulty. Such gain-oriented proxy signals could help more accurately identify frames that truly benefit from reprocessing by the strong detector and narrow the gap to oracle routing.

- **Extension to broader adaptive inference scenarios.** The current routing setting mainly focuses on the selection between a weak detector and a strong detector. Future work could extend the proposed difficulty estimation method to multi-model routing, dynamic resource allocation and real-time online deployment on edge devices.

6.4. Summary

Overall, this thesis demonstrates that temporal consistency can serve as a lightweight and interpretable signal for frame-level difficulty estimation and adaptive video object detection. Although the current method still has limitations in generalization ability, parameter sensitivity and score aggregation, the results indicate that proxy-based temporal consistency is a promising direction for efficient adaptive inference in video object detection systems.

A. Appendix

A.1. Why Ranking Quality Does Not Necessarily Equal Routing Gain

Intuitively, one may expect that an estimator achieving high ranking accuracy should also bring a large routing gain. Since adaptive routing prioritizes frames according to their estimated difficulty, a better ranking of difficult frames is likely to help with more reasonable detector allocation and thus improve the overall detection performance. However, ranking quality is not the same as routing effectiveness.

Ranking-oriented evaluation is used to judge whether the estimator can correctly distinguish relatively difficult frames from easier frames. Metrics such as pairwise accuracy, Kendall Tau coefficient, Pearson correlation coefficient and Spearman correlation coefficient are used to measure the consistency between the estimated difficulty ranking and the observed frame-level detection performance. However, routing-oriented evaluation introduces an additional factor: the performance gap between the weak detector and the strong detector.

Even if a frame is correctly identified as difficult, re-evaluation with the strong detector may not significantly improve detection performance. Some difficult frames remain challenging for both detectors, while some relatively easy frames may still benefit from stronger feature representations. Therefore, routing performance depends not only on ranking quality, but also on the performance improvement that the strong detector can provide.

In order to better understand this relationship, the correlation between frame-level weak-detector quality and normalized routing gain is analyzed. The raw routing gain is defined as the improvement obtained after replacing the weak-detector prediction with the corresponding strong-detector prediction. For each frame,

$$G_i^{\text{raw}} = AP_i^{\text{strong}} - AP_i^{\text{weak}},$$

where AP_i^{weak} and AP_i^{strong} denote the frame-level detection performance of the weak and strong detectors, respectively.

For visualization and correlation analysis, the gain is normalized by the maximum absolute gain over all evaluated frames:

$$G_i^{\text{norm}} = \frac{G_i^{\text{raw}}}{\max_k |G_k^{\text{raw}}| + \epsilon}.$$

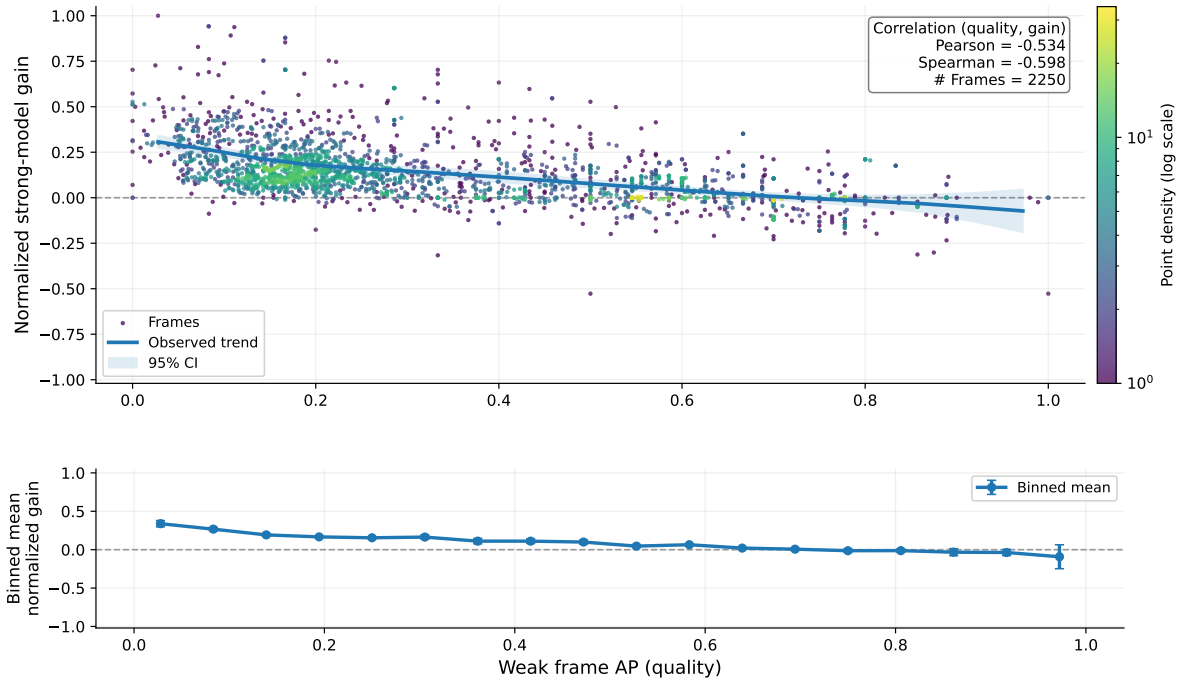


Figure A.1.: Relationship between frame-level weak-detector quality and normalized routing gain on MOT17. Lower-quality frames tend to exhibit larger potential gains on average, although the correlation is only moderate.

Figure A.1 shows the relationship between frame-level weak-detector quality and normalized routing gain. The upper part presents a frame-level scatter plot, while the lower part shows the corresponding binned average routing gain. The overall trend shows that low-quality frames can usually obtain larger benefits from strong-detector re-evaluation. This observation supports the routing hypothesis adopted in this thesis: frames estimated as difficult are more likely to bring greater performance improvements when processed by a stronger detector.

However, this relationship is far from deterministic. Although a moderate negative correlation can be observed, with a Pearson correlation coefficient of -0.534 and a Spearman correlation coefficient of -0.598, the variation across different frames is still significant. Many low-quality frames bring only limited improvements, while some high-quality frames can still achieve noticeable gains. Therefore, frame-level quality should be regarded as an informative proxy signal for routing gain, rather than a direct predictor of the gain itself.

The lower part of the figure further shows that, as frame-level quality improves, the average routing gain usually decreases. However, this trend is relatively flat and contains large fluctuations, indicating that quality alone cannot fully explain the gain obtained from strong-detector re-evaluation.

This observation helps explain why the improvement of ranking metrics does not necessarily lead to a corresponding improvement in routing performance. Routing effectiveness is determined by two factors: the ability of the estimator to identify difficult frames and the

performance improvement that the strong detector can actually provide. Therefore, even if an estimator has strong ranking quality, its routing performance can still be limited by the characteristics of the weak-strong detector pair.

Overall, ranking quality provides valuable information for adaptive detector routing, but routing performance is also limited by the achievable gain from strong-detector reprocessing. Therefore, even a highly accurate ranking strategy cannot recover gains that are not available for the selected weak-strong detector pair. This explains why the gap between actual routing strategies and the Oracle routing upper bound can remain large, even when the ranking accuracy is relatively high.

List of Figures

List of Tables

Bibliography

[1] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. "The pascal visual object classes (voc) challenge". In: *International journal of computer vision* 88.2 (2010), pp. 303–338.

[2] S. Ren, K. He, R. Girshick, and J. Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *Advances in neural information processing systems* 28 (2015).

[3] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. "End-to-end object detection with transformers". In: *European conference on computer vision*. Springer. 2020, pp. 213–229.

[4] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei. "Deep Feature Flow for Video Recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. July 2017.

[5] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. "Feature pyramid networks for object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2117–2125.

[6] Z. Cai and N. Vasconcelos. "Cascade r-cnn: Delving into high quality object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 6154–6162.

[7] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai. "Deformable detr: Deformable transformers for end-to-end object detection". In: *arXiv preprint arXiv:2010.04159* (2020).

[8] M. S. Ramanagopal, C. Anderson, R. Vasudevan, and M. Johnson-Roberson. "Failing to learn: Autonomously identifying perception failures for self-driving cars". In: *IEEE Robotics and Automation Letters* 3.4 (2018), pp. 3860–3867.

[9] Z. Cao, Z. Li, Y. Chen, H. Pan, Y. Hu, and J. Liu. "Edge-cloud collaborated object detection via difficult-case discriminator". In: *2023 IEEE 43rd International Conference on Distributed Computing Systems (ICDCS)*. IEEE. 2023, pp. 259–270.

[10] J. Qiu, R. Wang, B. Hu, R. Guérin, and C. Lu. "Optimizing edge offloading decisions for object detection". In: *2024 IEEE/ACM Symposium on Edge Computing (SEC)*. IEEE. 2024, pp. 164–177.

[11] X. Lu, Y. Cao, S. Liu, C. Long, Z. Chen, X. Zhou, Y. Yang, and C. Xiao. "Video Shadow Detection via Spatio-Temporal Interpolation Consistency Training". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 3116–3125.

[12] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang. "Acquisition of localization confidence for accurate object detection". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 784–799.

[13] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang. "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection". In: *Advances in neural information processing systems* 33 (2020), pp. 21002–21012.

[14] Y. Geifman and R. El-Yaniv. "Selective classification for deep neural networks". In: *Advances in neural information processing systems* 30 (2017).

[15] Y. Geifman and R. El-Yaniv. "Selectivenet: A deep neural network with an integrated reject option". In: *International conference on machine learning*. PMLR. 2019, pp. 2151–2159.

[16] D. Madras, T. Pitassi, and R. Zemel. "Predict responsibly: improving fairness and accuracy by learning to defer". In: *Advances in neural information processing systems* 31 (2018).

[17] H. Mozannar and D. Sontag. "Consistent estimators for learning to defer to an expert". In: *International conference on machine learning*. PMLR. 2020, pp. 7076–7087.

[18] W. Geng, N. Mohan, and J. Ott. "Budget-Adaptive Routing: Skipping the Weak When the Strong Answers Anyway". In: *Proceedings of the Workshop on Networks for AI Computing (NAIC '26)*. 2026. DOI: 10.1145/3789240.3828740.

[19] W. Geng, X. Su, N. Mohan, J. Ott, and P. Hui. "SMOOTH: Scalable Multitask Offloading with Backbone Sharing". In: *2026 IFIP Networking Conference (IFIP Networking)*. Lugano, Switzerland, May 2026, pp. 1–10.

[20] K. Bernardin and R. Stiefelhagen. "Evaluating multiple object tracking performance: the clear mot metrics". In: *EURASIP Journal on Image and Video Processing* 2008.1 (2008), p. 246309.

[21] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. "Performance measures and a data set for multi-target, multi-camera tracking". In: *European conference on computer vision*. Springer. 2016, pp. 17–35.

[22] J. Luiten, A. Osep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixé, and B. Leibe. "Hota: A higher order metric for evaluating multi-object tracking". In: *International journal of computer vision* 129.2 (2021), pp. 548–578.

[23] A. Milan, L. Leal-Taixé, I. Reid, S. Roth, and K. Schindler. "MOT16: A benchmark for multi-object tracking". In: *arXiv preprint arXiv:1603.00831* (2016).