Faculté de géographie et d'aménagement
Université de Strasbourg

MASTER 2

Observation de la Terre et Géomatique

2025 – 2026



Estimation de la masse de plastique en transit sur les cours d'eau : amélioration d'un pipeline de détection et de suivi embarqué

Morgane Carmier

Septembre 2026

Structure d'accueil : IPHC, Institut Pluridisciplinaire Hubert Curien
ZAEU, Zone Atelier Environnementale Urbaine
LIVE, Laboratoire Image, Ville, Environnement
UMR7362 CNRS-Université de Strasbourg
3, rue de l'Argonne F-67000 Strasbourg

Maîtres de stage : Gauthier Grimmer & Romain Wenger & Valentin Chardon
Laboratoire Image, Ville, Environnement
UMR7362 CNRS-Université de Strasbourg
3, rue de l'Argonne F-67000 Strasbourg

Laboratoire image, ville et environnement | LIVE
Université de Strasbourg & cnrs & ENGEES

Remerciements

Je tiens à remercier toutes les personnes qui ont contribué à la réussite de mon stage et qui m'ont aidée lors de la rédaction de ce mémoire.

Je souhaite tout d'abord remercier mon directeur de mémoire, M. Grimmer, doctorant à l'université de Strasbourg et responsable du projet KDP dans lequel s'inscrit ce travail, pour sa patience, sa disponibilité et ses conseils. Je vous remercie sincèrement pour m'avoir fait confiance sur ce sujet, merci d'avoir pris le temps de relire et critiquer ce mémoire.

Je remercie également l'équipe de l'université et du laboratoire LIVE avec laquelle j'ai pu travailler, et en particulier M. Wenger, pour ses conseils ainsi que pour avoir assuré la partie théorique de ma formation sur l'intelligence artificielle.

Au-delà de cela, je voudrais remercier mes professeurs, et notamment mon jury composé de Mme Puissant et de M. Herrault, pour la qualité des cours auxquels il m'a été permis d'assister au sein de l'université.

Merci aux différents acteurs du territoire qui ont rendu ce projet possible.

Je tiens ensuite à témoigner ma reconnaissance aux personnes suivantes, pour leur aide dans la réalisation de ce mémoire : M. Octave Blandet, qui m'a beaucoup aidée à développer ma réflexion en partageant ses connaissances et son expérience de l'intelligence artificielle. Merci d'avoir écouté mes divagations sur l'entraînement des modèles ainsi que sur l'estimation de la masse.

Mes camarades, pour leur soutien et leur entraide tout au long de l'année.

Ma famille, pour leur soutien constant et leurs encouragements, merci d'avoir pris le temps de relire ce travail.

Table des matières

Table des matières **5**

Table des figures **6**

Liste des tableaux **7**

1 Introduction **1**

2 État de l'art **5**
2.1 Chronologie de la détection plastique : 5
2.1.1 Méthodes traditionnelles de surveillance 5
2.1.2 De la surveillance manuelle à l'intelligence artificielle 6
2.1.3 Verrous scientifiques actuels 7
2.2 Les CNN : réseaux de neurones à convolution : 8
2.2.1 Cadre de l'IA et des CNN 8
2.2.2 Architecture des réseaux convolutifs 9
2.2.3 Détection d'objets et Edge AI 10
2.3 Vers le déploiement d'un modèle : 11
2.3.1 Fiabilité du modèle 11
2.3.2 Adaptation des modèles aux conditions réelles 12
2.3.3 Déploiement opérationnel 13
2.3.4 Estimation de masse 14

3 Méthodologie **17**
3.1 Aire d'étude 17
3.2 Données 18
3.2.1 Données dans la littérature 19
3.2.2 Création des nouveaux jeux de données 20
3.2.3 Choix de la résolution des images 24
3.3 Entraînement des modèles 29
3.3.1 Méthode Statistique 30
3.3.2 Méthode du K-fold 32
3.3.3 Méthode temporelle 35
3.4 Chaîne de traitement 37
3.4.1 Embarqué 37
3.4.2 Tracking 38
3.5 Estimation des dimensions réelles 39
3.5.1 Module d'estimation 3D 40
3.5.2 Application de correctif 43
3.5.3 Estimation du volume et de la masse 43
3.5.4 Sécurité géométrique et score de confiance 44

4 Résultats **47**
4.1 Entraînement et sélection des modèles 47
4.1.1 Impact de la résolution 47
4.1.2 Comparaison de l'échelle du réseau (YOLO26n/s/m) 47
4.1.3 Méthode statistique 48

4.1.4 Généralisation spatiale (validation croisée par site) 48
4.1.5 Entraînement par disposition temporelle 49
4.2 Déploiement embarqué 52
4.3 Suivi multi-objets 52
4.4 Estimation des dimensions et de la masse 53
4.4.1 Limites de la correction polynomiale de Grimmer 53
4.4.2 Validation sur données de suivi réelles 54
4.4.3 Estimation du volume et de la masse 54

5 Discussion **57**
5.1 Mise en perspective des résultats 57
5.1.1 Choix de l'architecture et performances 57
5.1.2 Méthode statistique 58
5.1.3 Généralisation spatiale et fuite de données 58
5.1.4 Effet du split temporel 59
5.1.5 Absence d'effet de la pondération de la loss 59
5.1.6 Embarqué 59
5.1.7 Suivi multi-objets 60
5.1.8 Robustesse de la correction géométrique de Grimmer et al., 2025 . 60
5.2 Limites méthodologiques 61
5.2.1 Représentativité des données d'acquisition 61
5.2.2 Biais et erreurs sur l'annotation 62
5.2.3 Limites de l'estimation dimensionnelle et de masse 62
5.2.4 Limites de la validation sur données réelles 63
5.3 Perspectives d'amélioration 64

6 Conclusion **66**

7 Annexe **74**
7.1 Déclaration d'utilisation de l'intelligence artificielle 74
7.2 Données dans la littérature 75
7.3 Glossaire 76

Table des figures

1 Aire d'étude, Eurométropole de strasbourg 17
2 Différents sites de l'étude, respectivement : Wolxheim , Holtzheim, Steingiessen amont et Steingiessen aval 20
3 Sites de nuit, respectivement : Holtzheim et Steingiessen amont 20
4 Classe d'objets observés (Bois,plastique, kayak) de nuit et de jour 21
5 distribution des dimensions des annotations 25
6 Principe général de la validation croisée par K-fold. 32
7 méthode temporelle 35
8 séquence d'observation des annotation sur les vidéo 36
9 Chaîne de traitement complète de l'estimation de la masse, depuis la détection YOLO en sortie du trackeur et le suivi jusqu'à l'application du facteur correctif final. 40
10 Précision et rappel moyens (± écart-type) obtenus par site de test lors de la validation croisée par site, modèle YOLO26s. 49

11 Résultats de masse par piste, Field 1. 54
12 Résultats de masse par piste, Field 2. 55

Liste des tableaux

1 Composition de la donnée de jour. 21
2 Composition de la donnée de nuit. 21
3 Typologie des objets à détecter. Cette table détaille la typologie réfléchie lors des travaux déjà effectués sur cette zone d'étude notamment par Grimmer et al., 2025 22
4 Éléments les plus petits 26
5 Calibration des résolutions de caméras 29
6 Organisation comparée des jeux de données D1 (split par site, Grimmer et al.) et D2 (split aléatoire 80/20, notre étude). n.c. : proportion d'items non communiquée pour ce split. 31
7 Configurations de la validation croisée par site : pour chaque combinaison, l'entraînement utilise les trois sites restants. Sam : Steingiessen amont ; Sav : Steingiessen aval ; H3 : Holtzheim mars ; H4 : Holtzheim avril ; W3 : Wolxheim mars. 34
8 Impact de la résolution d'image d'entrée sur les performances de test de YOLO26s (GPU NVIDIA RTX 4000 Ada). 47
9 Comparaison des performances de test selon la taille du réseau YOLO26 (GPU NVIDIA RTX 4000 Ada). 48
10 Comparaison des Configurations 1 et 2 (YOLOv8n, §3.3.1) à la référence publiée par Grimmer et al., 2025. [a] Rappel disponible par classe uniquement (0,90/0,99/0,98 pour la Config. 1 ; 0,84/n.d./0,00 pour la Config. 2). [b] Classe absente du jeu de validation retenu pour cette configuration. [c] mAP50 calculé sur les classes 1 et 3 uniquement ; non directement comparable aux autres valeurs de la colonne. 48
11 Synthèse de la validation croisée par site : effectifs de test et performance (moyenne ± écart-type sur les 4 combinaisons d'entraînement testées, n=4 par site). 49
12 Comparaison des modèles de nuit et mixte 50
13 Matrices de confusion normalisées par condition d'acquisition (jour, nuit, mixte) / Courbe précision-rappel conditions diurnes, nocturne et mixte . . 51
14 Stabilité métrique et latences comparées après conversion (banc de test : PC de laboratoire). 52
15 Performance sur cible embarquée (Raspberry Pi 5, résolution d'entrée 896×512 px) des trois formats d'inférence pour le modèle YOLO26s retenu en production. 52
16 Comparaison des trackers : nombre de pistes stables (au moins 8 frames) au regard du nombre réel d'objets introduits par séquence. 53
17 Comparaison globale sur les 2048 détections appariées : pipeline de Grimmer appliqué sans garde-fou vs notre adaptation. 53
18 Erreur absolue moyenne par catégorie d'objet : pipeline de Grimmer sans garde-fou vs notre adaptation (garde-fous, extrapolation linéaire, branche bidon dédiée). 53

1 Introduction

Les écosystèmes aquatiques recouvrent environ 70 % de la surface terrestre, et la préservation de leur intégrité constitue l'un des enjeux majeurs du XXIe siècle. Pourtant, ces espaces vitaux sont aujourd'hui gravement menacés par l'introduction massive de débris anthropiques, qui ont modifié irrémédiablement les équilibres naturels. Contrairement aux débris organiques qui se décomposent rapidement, les plastiques sont conçus pour être durables et résistants, ce qui entraîne une persistance dans l'environnement pouvant s'étendre sur plusieurs décennies, voire des siècles (van Emmerik & Schwarz, 2020). Cette longévité, couplée au faible coût de production, est à l'origine d'une fabrication mondiale massive qui a remplacé peu à peu des matériaux plus lourds comme le verre ou l'acier, favorisant leur accumulation dans l'ensemble des milieux.

Une fois rejetés dans l'environnement, les plastiques ne disparaissent pas, ils subissent une lente fragmentation. Sous l'effet des rayonnements UV, de l'action mécanique de l'eau ainsi que des variations thermiques, ils se décomposent progressivement en résidus de tailles décroissantes, (Galgani et al., 2019 ; Thompson et al., 2004). Les résidus qui s'en détachent peuvent atteindre différents stades. On distingue alors les macroplastiques, objets ou fragments supérieurs à 5 mm, les microplastiques qui composent la partie visible inférieure à 5 mm et enfin les nanoplastiques, dont les dimensions peuvent descendre en dessous du micromètre (Lebreton et al., 2024 ; Tramoy, Gasperi, Colasse & Tassin, 2020). En 2025, les recherches de Hossain et Engelhardt, 2025 ont mis en exergue une « empreinte plastique globale » de laquelle il découle que les microplastiques et nanoplastiques sont une menace pour la totalité des écosystèmes. La taille infime de ses particules les rend particulièrement difficiles à caractériser et à surveiller : plus elles sont petites, plus elles se dispersent largement dans l'environnement et plus leur détection requiert des méthodes sophistiquées.

Depuis 1990, des études comme le rapport du GESAMP établissent un constat alarmant en estimant qu'environ 80 % de la pollution marine provient de sources terrestres. Ce constat révèle que la protection des mers se joue bien en amont des océans : les rivières agissent en effet comme des vecteurs facilitant le transport de pollution en reliant les zones urbaines et industrielles aux milieux marins océaniques (van Emmerik & Schwarz, 2020). Cette pollution y est alors exacerbée par la dynamique des courants marins et l'intensité du transport maritime international (Luz et al., 2025), transformant une problématique locale en un défi d'ordre mondial. L'urgence est d'autant plus marquante selon les projections de Lebreton et Andrady, 2019. Sans intervention, la génération des déchets plastiques devrait s'aggraver exponentiellement d'ici 2050, faisant de l'étude des bassins versants et des rivières un terrain d'étude et d'action prioritaire dans la lutte contre la pollution des espaces marins.

Les conséquences qu'engendrent les débris plastiques sur l'échelle du vivant sont documentées à toutes les échelles. Les travaux de Barnes et al., 2009 ayant démontré que l'accumulation des débris fragmentés dépasse largement la simple pollution de surface, en affectant l'ensemble des espaces hydriques. L'altération de la santé des océans, soulignée par Galgani et al., 2019, engendre des impacts écosystémiques globaux perturbant l'équilibre biologique des milieux marins. Les conséquences sur la faune sont dévastatrices et se classent en deux catégories principales (van Emmerik & Schwarz, 2020) : l'ingestion des résidus fins et l'enchevêtrement des gros éléments, tels que les bouteilles ou les filets de pêche abandonnés. L'ingestion concerne des centaines d'espèces, des tortues aux poissons, provoquant des obstructions intestinales, une réduction de la condition physique

et une altération des processus de reproduction. L'enchevêtrement reste, quant à lui, une cause fréquente de mortalité chez les espèces marines et aviaires.

Enfin, ces débris impactent directement l'activité humaine. Ils agissent comme des vecteurs de polluants chimiques toxiques et persistants qui, par le biais de la bioaccumulation (Thompson et al., 2004), remontent la chaîne alimentaire jusqu'à menacer la santé humaine. Au-delà de l'aspect sanitaire, ces déchets perturbent l'économie en causant des pertes dans les secteurs du tourisme et de la navigation. En milieu urbain, l'accumulation de plastiques dans les infrastructures de drainage augmente significativement les risques d'inondation en obstruant le passage des eaux, aggravant ainsi les risques naturels pour les populations (Lebreton et al., 2024 ; van Emmerik & Schwarz, 2020)

La mise en avant des risques que le plastique peut faire encourir à l'écosystème marin ainsi qu'à l'homme va de pair avec l'augmentation des travaux s'intéressant à sa détection ainsi qu'à l'estimation des fuites de plastique vers l'océan.

Parallèlement, des acteurs de la société civile comme l'ONG Surfrider ou Naturaul'un partenaires des travaux de Tramoy, Gasperi, Colasse et Tassin, 2020, apportent une expertise de terrain indispensable, complémentaire aux outils de modélisation développés. Dans le secteur privé, des solutions telles que des entreprises comme Noria Innovateurs durables, s. d. ou Obscape, s. d. déploient des technologies de plus en plus sophistiquées, en accompagnant la détection de systèmes de captures des débris ainsi que des dispositifs de surveillance de la tranche humide.

Malgré la dynamique multi-acteurs qui se déploie, l'analyse de l'état de l'art révèle des lacunes persistantes qui freinent encore l'efficacité des stratégies de remédiation.

En effet le développement et le déploiement de solutions créent des coûts qui pèsent sur les institutions, notamment publiques, visées par ces produits. Le matériel nécessaire à une surveillance en continue est coûteux en ressources, car nécessitant des installations qui leur soient propres, mais aussi en main d'œuvre pour leur entretien (Noria Innovateurs durables, s. d. ; Obscape, s. d.). On constate par conséquent dans la plupart des solutions à bas coût une absence de détections en continu, comme chez Sio et al., 2022 ce qui crée un biais temporel, la possibilité d'avoir des cameras 24 h/24 sur site peut permettre de détecter de potentielles modifications des dynamiques de transport avec des évènements majeurs lors de périodes de crue ou de forte intempérie. Pourtant même si des travaux comme ceux de Grimmer et al., 2025 existent avec une observation en temps réel du cours d'eau via des cameras in-situ ils n'incluent pas de solutions pour la surveillance nocturne où de tels évènements sont aussi susceptibles d'arriver et où les cameras RGB habituelles sont aveugles.

Ces observations laissent donc place à une dernière question, celle de la masse en transit sur les rivières. La masse de plastique en transit reste souvent une inconnue car bien souvent on se contente du comptage van Emmerik et Schwarz, 2020 ou de l'estimation via la projection de l'objet sur l'image (Grimmer et al., 2025). Or, cette donnée permet de dimensionner les installations de récupération et de garantir leur efficacité (Noria Innovateurs durables, s. d.). La quantification ainsi que l'estimation de la masse de débris charriés peuvent aussi être des statistiques à transmettre aux gestionnaires de milieux naturels afin d'organiser des campagnes de collecte régulières.

Pour pallier les principaux manques qui ont été relevés dans la littérature, la mise en place d'un système de vision par ordinateur pour la détection d'objets a été pensée. En effet, un tel système répondrait aux principaux soucis de gestion que nous avons pu observer. Un appareil périphérique conçu de tel manière permettrait une surveillance constante des transports de débris anthropiques par les rivières en amont et en aval de l'Ill.

Dans quelle mesure des modèles de détection d'objets basés sur l'apprentissage profond peuvent-ils permettre une surveillance continue, robuste et à faible coût des déchets en rivière sur des dispositifs embarqués à ressources limitées ?

Afin de répondre à cette problématique nous allons nous baser sur plusieurs hypothèses de recherche. Premièrement nous supposerons que l'augmentation de la taille des dataset utilisés va augmenter la qualité des modèles. En comparant nos données à un modèle de référence on va essayer d'améliorer ses capacités : ici avec une comparaison aux modèles de Grimmer et al., 2025

Deuxièmement, nous faisons l'hypothèse que l'entraînement d'un modèle capable de détecter les débris en rivière de nuit comme de jour permettrait de limiter la taille sur la carte mère. Les limites de performance de l'Edge computing, ainsi que les besoins de l'étude d'une observation des sites en temps réel sont à prendre en considération lors de la conception.

Troisièmement, nous supposons que l'intégration de données acquises en conditions de faible luminosité contribue à améliorer la détection des déchets durant les périodes nocturnes.

Enfin, nous faisons l'hypothèse que les caractéristiques géométriques extraites des objets détectés permettent d'estimer leur taille et leur volume avec une précision suffisante pour produire des indicateurs quantitatifs du transport de déchets plastiques en rivière.

Ce travail s'articule autour de plusieurs objectifs. Il s'agit, dans un premier temps, de constituer des jeux de données multimodaux (RVB et infrarouge) adaptés à l'entraînement de modèles de détection. Dans un second temps, différentes architectures de réseaux de neurones de la famille YOLO, seront évaluées et optimisées afin de répondre aux contraintes de l'embarqué. Enfin, une méthode de quantification des débris sera proposée, basée sur l'estimation de leur taille et de leur volume à partir des observations visuelles.

Afin de tester nos hypothèses, nous allons reposer notre méthodologie sur une approche expérimentale en combinant différents sets de données à partir de différents sites d'étude, mais aussi en entraînant différents modèles avec différents paramètres d'optimisation. Nous réentraînerons nos modèles optimisés avec de nouvelles données de nuit en vue d'obtenir un modèle ambivalent et léger pouvant être embarqué sur des appareils à ressources limitées. Puis, à l'aide de la régression linéaire pensée par Grimmer (2025), nous estimerons la taille des plastiques détectés par le modèle et, en les comparant à des valeurs réelles, nous testerons l'efficacité de la méthode.

Notre réflexion se déroulera donc en plusieurs points. Dans un premier temps, un état de l'art présentera les travaux existants en détection de débris aquatiques et en vision embarquée. Ensuite, la méthodologie de constitution des jeux de données et l'entraînement des modèles sera détaillée avec l'entraînement du modèle YOLO, «You Only Look Once», à faible coût énergétique (version nano ainsi que technique de diminution de la taille des données). Les résultats expérimentaux seront ensuite analysés, afin de proposer une discussion sur les limites et les perspectives d'amélioration du système développé.

2 État de l'art

2.1 Chronologie de la détection plastique :

2.1.1 Méthodes traditionnelles de surveillance

Comme présenté en introduction, les rivières constituent les principaux vecteurs de transfert de déchets plastiques vers les océans (Lebreton et al., 2024). Leur détection représente depuis plusieurs décennies un enjeu majeur pour la protection de l'environnement. Historiquement, la surveillance des plastiques reposait sur des méthodes de comptage manuelles, telles que les observations de macroplastiques depuis des ponts ou des berges comme réalisées par van Emmerik et al., 2018, ainsi qu'avec des prélèvements par filets (« manta trawls ») développés par Moore et al., 2011. Ces approches ont permis d'établir les premiers bilans mondiaux et les premières estimations de la pollution plastique, mais ses méthodes restent limitées car nécessitant de fortes mobilisations de main-d'œuvre sans pouvoir assurer de suivi en continu, avec des biais liés aux conditions environnementales (courant, débit, flux de débris) et un facteur extérieur pouvant déconcentrer le chercheur (van Emmerik et al., 2018). Au fil des travaux effectués, des constats ont accentué l'importance des données, ceux de Lebreton et al., 2024 ; Tramoy, Gasperi, Colasse et Tassin, 2020 et ont permis de mettre en évidence l'augmentation importante de la fragmentation des plastiques dans les océans, avec des concentrations pouvant être multipliées par cinq en sept ans dans certaines zones.

Face aux limites des approches de terrain, plusieurs solutions technologiques ont été développées afin d'améliorer le suivi des déchets flottants. Parmi elles, la surveillance par RFID (identification par radiofréquence) proposée par Chardon et al., 2025 qui permet d'étudier la mobilité des macro-débris et d'identifier des zones de piégeage, comme les embâcles, contribuant à une meilleure compréhension de leur transport et de leur fragmentation en amont des milieux marins (Tramoy, Gasperi, Nguyen & Tassin, 2020). Toutefois, cette méthode ne fournit pas d'information sur la quantité de déchets transportés ni sur leur nature.

En complément, la télédétection s'est progressivement imposée comme une solution adaptée à différentes échelles spatiales. Les données satellitaires, notamment issues du programme Copernicus de l'Agence spatiale européenne, permettent de détecter les accumulations de macrodéchets en haute mer (Lavender, 2022). À une échelle spatiale intermédiaire, les drones facilitent l'acquisition d'images dans des secteurs difficilement accessibles et contribuent à la constitution de bases de données environnementales (Das et al., 2026). Cependant, ces campagnes d'instrumentation sont coûteuses et ponctuelles se qui constitue un frein pour la surveillance en permanence des cours d'eaux.

Afin d'assurer un suivi continu, plusieurs études, notamment celle de Sio et al., 2022, proposent l'installation de caméras fixes sur des infrastructures telles que des ponts. Ces dispositifs enregistrent en continu la surface des cours d'eau et s'affranchissent ainsi des limites liées aux observations ponctuelles. Ils offrent la possibilité de suivre l'évolution des flux de déchets en temps réel, mais génèrent un volume considérable de données (« Big Data »), rendant leur exploitation manuelle difficile.

Cette production continue d'images a favorisé le développement de méthodes automatisées reposant sur la vision par ordinateur et l'intelligence artificielle, capables de traiter efficacement ces grands volumes de données et d'assurer une surveillance continue des déchets plastiques.

2.1.2 De la surveillance manuelle à l'intelligence artificielle

Les progrès de la vision par ordinateur ont permis l'intégration de l'intelligence artificielle (IA) aux dispositifs de détection existants afin d'automatiser le traitement des observations et de proposer des méthodes de suivi reproductibles Sio et al., 2022 ; Tamin et al., 2022. Cette automatisation répond directement à la difficulté du traitement en continu de grands volumes de données (« Big Data »), tout en limitant les biais associés aux actions humaines.

Le domaine médical, précurseur dans l'utilisation de la vision par ordinateur, constitue une référence pour l'évaluation des performances de ces méthodes. Les travaux de Salinas et al., 1997 montrent que les systèmes automatisés réduisent fortement les erreurs de comptage par rapport aux méthodes manuelles, avec des coefficients de variations passant de 27,9 % à 2,9 % sur des densités élevées. Plus récemment, Perez-Torrents et Minvielle, s. d. montre que les performances des modèles d'IA deviennent comparables à celles des radiologues, avec un taux d'erreur de 7,5 % contre 3,5 % pour l'expert humain, tandis que la combinaison des analyses homme-machine réduit l'erreur à seulement 0,5 %. Ces résultats illustrent la capacité de l'IA à assurer une surveillance continue, 24 h/24 et 7 j/7, sans les limitations liées à la fatigue humaine.

Dans le domaine de la détection des déchets plastiques, le Deep Learning est aujourd'hui largement utilisé à travers différentes architectures de vision par ordinateur (**Soumya2023** ; Sio et al., 2022). Sur les chaînes de tri, ces modèles offre la possibilité de localiser automatiquement les déchets et d'identifier leur nature (PET, PE, etc.) à partir d'images ou de vidéos, avec des précisions pouvant atteindre 95 %. Toutefois, ces applications reposent généralement sur des conditions d'acquisition contrôlées, avec des caméras placées à la verticale et des arrière-plans homogènes, très différents des conditions rencontrées en rivière. Malgré ces contraintes, ces travaux démontrent le potentiel du Deep Learning pour automatiser la surveillance des déchets flottants.

Les premières approches de vision par ordinateur étaient principalement limitées à la classification d'images, permettant uniquement de déterminer la présence globale de plastique dans une image. Pour affiner le suivi environnemental les techniques ont progressivement évolué vers la segmentation, capable d'isoler précisément les contours des déchets, puis vers la détection d'objets, qui combine simultanément leur localisation et leur identification.

Cette évolution est illustrée par les travaux de Grimmer et al., 2025, qui a développé un pipeline complet combinant la géométrie et le Deep Learning afin de détecter en continu les déchets flottants à l'aide de caméras fixes installées in situ. Son application se base sur des architectures déjà évaluées par Sio et al., 2022 sur des modèles de même nature pour la détection de plastique en rivière. Cette approche constitue l'une des premières démonstrations d'un système opérationnel capable d'assurer une surveillance autonome et continue.

Malgré ces avancées, les systèmes existants restent fortement dépendants des conditions d'acquisition. Les variations de luminosité, les conditions météorologiques dégradées ou les observations nocturnes limitent encore leur robustesse, alors que les principaux flux de déchets sont souvent associés aux épisodes de crue. Ces limites montrent que, malgré les progrès récents de la détection automatisée, plusieurs verrous scientifiques et techniques restent à lever avant le déploiement de systèmes de surveillance réellement continus en rivière.

2.1.3 Verrous scientifiques actuels

Un premier défi concerne l'identification des déchets dans l'environnement, les plastiques possèdent des signatures spectrales spécifiques, notamment dans le proche infrarouge (NIR), qui rend leur détection à grande échelle possible. Des indices optiques tels que le Floating Debris Index (FDI) exploitent ainsi les données satellitaires du programme Copernicus pour identifier les accumulations de déchets flottants (Lavender, 2022). Toutefois, leur application aux cours d'eau de faible largeur reste limitée par la résolution spatiale des satellites et les perturbations liées à l'environnement riverain. À plus petite échelle, les travaux de Martínez-Hernández et al., 2024 montrent que différents polymères présentent des pics de réflectance caractéristiques, notamment autour de la longueur d'onde $\lambda = 833$ nm, permettant de distinguer les plastiques des éléments naturels dans des conditions favorables.

Les systèmes de détection locaux reposent généralement sur des capteurs spécialisés ou des caméras à haute résolution (Noria Innovateurs durables, s. d. ; Obscape, s. d.). Bien que performants, leur coût d'acquisition et de maintenance limite leur déploiement sur de vastes réseaux hydrographiques. Cette contrainte explique que peu de dispositifs soient aujourd'hui exploités de manière opérationnelle à grande échelle, malgré les nombreuses solutions recensées dans la littérature. Ainsi, Winterstetter et al., 2021 identifient 51 solutions de surveillance des déchets, différenciées selon leur domaine d'application et leur niveau d'automatisation, mais seules quelques-unes font l'objet d'un déploiement permanent.

Un autre verrou majeur concerne les jeux de données utilisés pour entraîner les modèles d'intelligence artificielle. Bien que les publications se soient multipliées depuis 2022, la plupart des travaux reposent sur des bases acquises dans un contexte géographique unique et pour des objectifs spécifiques. Cette hétérogénéité complique la comparaison des performances et limite la capacité des modèles à généraliser vers de nouveaux environnements. Les variations de couleur de l'eau, de turbidité, de végétation, de conditions météorologiques, d'angle de prise de vue ou de qualité des caméras modifient fortement l'apparence des déchets, conduisant souvent les modèles à apprendre des caractéristiques propres au site d'entraînement plutôt que celles des plastiques eux-mêmes.

Cette problématique est renforcée par différents phénomènes de fuite de données (« data leakage ») décrits par Apicella et al., 2025. Les fuites contextuelles, spatiales ou temporelles, notamment lorsque des images issues d'un même site ou d'une même séquence vidéo sont utilisées à la fois pour l'entraînement et la validation, conduisent à une surestimation artificielle des performances (Grimmer et al., 2025).

Ces limites mettent en évidence la nécessité de développer des approches capables de fonctionner dans des conditions variées à partir de jeux de données diversifiés et représentatifs de la diversité des contextes fluviaux. Malgré les progrès de la vision par ordinateur appliquée à la surveillance environnementale, peu de systèmes assurent aujourd'hui une détection automatisée, robuste et continue des déchets plastiques en conditions réelles. Ce constat justifie le développement d'une architecture de Deep Learning embarquée, capable d'être déployée directement sur site tout en maintenant un niveau de performance satisfaisant.

2.2 Les CNN : réseaux de neurones à convolution :

2.2.1 Cadre de l'IA et des CNN

L'intégration de l'intelligence artificielle (IA) dans la surveillance environnementale répond aux limites des méthodes traditionnelles, qui nécessitent une intervention humaine importante pour le traitement des données. Elle rend ainsi possible d'automatiser l'analyse de grands volumes de données (« Big Data ») de manière continue et reproductible (Alotaibi & Nassif, 2024). L'IA désigne l'ensemble des méthodes offrant à la machine la possibiliter d'exécuter des tâches nécessitant habituellement des capacités humaines, telles que la reconnaissance d'objets ou la prise de décision (Jordan & Mitchell, 2015). Parmi celles-ci, le Machine Learning permet à un modèle d'améliorer automatiquement ses performances à partir des données, sans reposer sur des règles prédéfinies, en apprenant des relations entre les caractéristiques d'entrée et les variables de sortie (Dey, 2016 ; Kumar, 2017 ; Naqa & Murphy, 2015).

Le Deep Learning constitue aujourd'hui l'une des principales approches du Machine Learning. Il repose sur des réseaux de neurones artificiels composés de couches consécutives capables d'apprendre automatiquement des représentations de plus en plus complexes de la donnée (LeCun et al., 2015). Parmi les différents paradigmes d'apprentissage, l'apprentissage supervisé est particulièrement adapté aux tâches de classification et de détection d'objets. Il s'appuie sur des jeux de données annotés permettant au modèle d'ajuster progressivement ses paramètres afin de réduire l'écart entre ses prédictions et les annotations de référence (Dey, 2016 ; Pereira et al., 2009). Les performances obtenues dépendent directement de la qualité et de la diversité de la donnée d'entraînement, qui oriente la capacité du modèle à généraliser à de nouvelles situations.

La vision par ordinateur (ou Computer Vision) constitue une branche de l'intelligence artificielle dédiée à l'analyse automatisée des images. Avant les années 1990, elle reposait principalement sur des filtres de convolution fixes et des règles prédéfinies par l'utilisateur pour extraire les caractéristiques telles que les contours, les textures ou les formes. L'apparition de réseaux de neurones convolutifs (CNN), avec des architectures fondatrices comme LeNet-5 par Lecun et al., 1998 puis AlexNet par **Krizhevsky2012**, a profondément transformé cette approche en permettant un apprentissage automatique de bout en bout (« end-to-end ») entièrement géré par ordinateur.

Les CNN reposent sur une succession de filtres et de regroupements (pooling) (**Schwindt2024** ; LeCun et al., 2015). Les couches de convolution appliquent des matrices afin d'extraire progressivement des caractéristiques visuelles telles que les contours, les textures ou les formes, tandis que les couches de pooling réduisent la dimension des données afin de diminuer le coût de calcul et d'améliorer la robustesse du modèle face aux variations de position des objets (Dey, 2016). Cette capacité d'apprentissage permet aujourd'hui d'obtenir des performances dépassant souvent celles des approches techniques manuelles (Alotaibi & Nassif, 2024 ; LeCun, 2016).

Toutefois, l'analyse d'images ne se limite pas à attribuer une étiquette à une scène. Les applications de surveillance environnementale nécessitent également de localiser précisément les objets détectés. Cette évolution a conduit au développement d'architectures spécialisées en détection d'objets, capables de combiner simultanément classification et localisation, ce qui constitue aujourd'hui la base des systèmes modernes de surveillance automatisée.

2.2.2 Architecture des réseaux convolutifs

L'architecture des réseaux de neurones convolutifs (CNN) étant constituée d'une succession de couches de traitement, leur organisation possède une importance. En effet, selon leur disposition, le réseau répondra à différents objectifs. Les modèles de classification, comme ResNet ou EfficientNet, attribuent une image à une classe à partir des caractéristiques extraites (Howard et al., 2017 ; LeCun et al., 2015). Mais il existe également des modèles de détection d'objets qui permettent d'aller plus loin en ajoutant à la classification la localisation des objets présents dans l'image.

Pour répondre à ces deux objectifs, les architectures modernes sont généralement organisées en trois parties. Le Backbone, est l'organe qui extrait les caractéristiques visuelles de l'image à l'aide d'un CNN pré-entraîné. Le Neck, fusionne les informations issues de différents niveaux du backbone afin d'améliorer la détection d'objets de tailles variées. Il fait le lien entre les deux étapes via une architecture spécifique (FPN ou PANnet) permettant d'extraire des caractéristiques à différentes échelles. Enfin vient le Head qui produit les prédictions finales en estimant la position des boîtes englobantes (*bounding boxes*), leur score de confiance ainsi que la classe de chaque objet (Lin et al., 2017 ; Liu et al., 2016).

Le choix d'une architecture dépend principalement du compromis entre précision et vitesse d'inférence. Les approches Two-Stage, telles que Faster R-CNN ou Mask R-CNN, génèrent d'abord des zones d'intérêt avant de les classifier. Elles offrent une excellente précision mais nécessitent des ressources de calcul importantes, limitant leur utilisation en temps réel. À l'inverse, les approches One-Stage, comme SSD ou RetinaNet, réalisent directement la détection en une seule étape, ce qui améliore considérablement la rapidité d'exécution (Lin et al., 2017 ; Liu et al., 2016 ; Ren et al., 2016).

Parmi ces architectures, la famille YOLO (You Only Look Once) s'est imposée comme une référence pour les applications temps réel grâce à son excellent compromis entre vitesse, précision et consommation de ressources (Redmon et al., 2016 ; Ultralytics, 2026). Les versions successives ont progressivement amélioré les performances de détection grâce à des évolutions telles que les têtes de détection découplées ou les approches Anchor-Free, tout en augmentant progressivement la complexité des calculs.

Afin de répondre aux contraintes des applications embarquées, YOLO26 (Ultralytics, 2026 Jocher et Qiu, 2026) introduit plusieurs optimisations destinées à réduire la latence d'inférence. Son architecture supprime notamment l'étape de Non-Maximum Suppression (NMS) grâce à un apprentissage de bout en bout, ce qui simplifie la tête de détection et améliore la convergence du modèle. Une autre stratégie mise en application est la STAL, qui renforce la détection des très petits objets, particulièrement fréquents dans les scènes naturelles.

Grâce à ces optimisations, YOLO26 est annoncé jusqu'à 43 % plus rapide sur processeur (CPU) que YOLO11 tout en conservant un niveau de précision élevé (Jocher et Qiu, 2026). Ces caractéristiques en font un modèle particulièrement adapté à la surveillance environnementale, où les conditions d'acquisition sont très variables (reflets, turbidité, météo ou faible luminosité). Associé à des caméras fixes ou à des drones, il est destiné à devenir une solution adaptée au suivi automatisé des déchets plastiques et à leur déploiement sur des dispositifs d'Edge AI (**The Ocean Cleanup** ; **2026** ; Bianchi & Putro, 2024 ; Grimmer et al., 2025).

2.2.3 Détection d'objets et Edge AI

L'Edge AI désigne l'exécution de modèles d'intelligence artificielle directement sur des dispositifs situés au plus près de la source de données, sans recourir à une infrastructure de calcul distante. Contrairement au Cloud Computing, où les images sont transmises vers un serveur pour être analysées, l'inférence est réalisée localement sur le système embarqué. Cette approche réduit les besoins en bande passante, améliore la confidentialité des données et permet au système de fonctionner même en l'absence de connexion réseau. Dans le domaine de la surveillance environnementale, où les caméras sont souvent installées sur des sites isolés et alimentées par des ressources limitées, l'Edge AI constitue une solution particulièrement adaptée (Shi et al., 2016).

Le choix du matériel conditionne directement les performances du système embarqué. Les plateformes basées sur des microcontrôleurs, telles que les cartes Arduino, sont adaptées aux applications TinyML et au traitement de capteurs simples, mais leurs capacités de calcul restent insuffisantes pour exécuter en temps réel des réseaux de neurones convolutifs de type YOLO sur des flux vidéo. À l'inverse, le Raspberry Pi 5 offre un compromis intéressant entre coût, consommation énergétique et performances. Équipé d'un processeur Broadcom BCM2712 et d'une mémoire vive pouvant atteindre 16 Go, il dispose des ressources nécessaires pour l'acquisition des images, leur prétraitement et l'exécution d'applications de vision par ordinateur. Ses performances peuvent être renforcées par un accélérateur dédié, tel que le Raspberry Pi AI Kit intégrant le processeur Hailo-8L, capable d'atteindre jusqu'à 13 TOPS, permettant ainsi l'exécution efficace de modèles comme YOLO26 tout en conservant une faible consommation énergétique.

Le déploiement d'un modèle sur un système embarqué nécessite toutefois une phase d'optimisation adaptée au matériel cible. Les moteurs d'inférence, tels que HailoRT, OpenVINO, TensorFlow Lite, ONNX Runtime ou NCNN, permettent d'exécuter efficacement les modèles en exploitant les capacités matérielles disponibles (Nguyen Tien et al., 2026). Leur utilisation implique généralement une conversion des poids obtenus après l'entraînement vers un format compatible avec le moteur choisi. Parmi ces formats, ONNX constitue aujourd'hui un standard largement adopté grâce à sa compatibilité avec de nombreux environnements d'exécution, tandis que d'autres solutions, comme TensorFlow Lite ou NCNN, reposent sur des formats spécifiques (Nguyen Tien et al., 2026 ; Ultralytics, 2026).

L'association d'architectures de détection performantes, comme YOLO26, avec des plateformes d'Edge AI permet ainsi de concevoir des systèmes capables d'assurer une détection automatisée directement sur le terrain, sans dépendre d'une connexion permanente à un serveur distant. Cette approche répond aux contraintes de la surveillance environnementale en combinant traitement en temps réel, faible consommation énergétique et autonomie de fonctionnement, ouvrant la voie au déploiement de réseaux de surveillance continus des déchets plastiques en rivière.

2.3 Vers le déploiement d'un modèle :

2.3.1 Fiabilité du modèle

Les performances d'un modèle de Deep Learning ne dépendent pas uniquement de son architecture, mais surtout de la qualité des données utilisées pour l'entraînement. Dans le contexte environnemental, assurer la fiabilité d'un modèle suppose de maîtriser l'ensemble de la chaîne de production des données, depuis leur acquisition jusqu'à leur exploitation pour produire des mesures quantitatives. Contrairement aux jeux de données standardisés de la vision par ordinateur, les images acquises en rivière sont soumises à de nombreuses perturbations, telles que les variations d'éclairement, les reflets à la surface de l'eau, la turbidité, les conditions météorologiques ou les fluctuations saisonnières, qui modifient l'apparence des objets (Lavender, 2022 ; Martínez-Hernández et al., 2024 ; Shukhratov et al., 2024). Cette variabilité complique à la fois la détection et la constitution de bases de données représentatives des conditions réelles.

La quantité de données disponibles ne garantit toutefois pas la fiabilité d'un modèle supervisé si les annotations sont incomplètes ou imprécises. Les réseaux de neurones apprennent directement à partir des informations fournies par l'étiquetage des images, de sorte que le bruit dans les labels limite leur capacité de généralisation (Minaee et al., 2022). Les outils d'annotation assistée et les plateformes collaboratives facilitent la création de jeux de données homogènes (Das et al., 2026 ; Ilani & Banadaki, 2025), tandis que le recours à plusieurs annotateurs ou à des procédures de validation par consensus permet de réduire les biais d'interprétation, notamment pour les catégories d'objets peu représentées. Même avec des annotations de qualité, les jeux de données couvrent rarement toute la diversité des situations rencontrées sur le terrain. Les techniques de *data augmentation* permettent ainsi d'accroître artificiellement cette diversité par des transformations telles que les rotations, les variations de luminosité ou de contraste, renforçant la robustesse des modèles (Shorten & Khoshgoftaar, 2019). Plus récemment, les modèles de diffusion offrent la possibilité de générer des images réalistes tout en préservant les annotations associées, ouvrant la voie à une augmentation plus représentative des situations sous-échantillonnées (Islam et al., 2024).

Une fois le modèle entraîné, ses performances doivent être évaluées afin de vérifier qu'elles traduisent une réelle capacité de généralisation et non une simple mémorisation des données d'entraînement. Les protocoles de validation croisée fournissent une estimation plus robuste des performances en limitant la dépendance à un unique découpage entraînement/test (Ramos et al., 2024 ; Shobaki & Milanova, 2025). Cette évaluation doit être complétée par des comparaisons avec d'autres approches sur des jeux de données et des métriques communes. Dans le domaine des déchets plastiques, cet exercice reste toutefois limité par l'absence de jeux de données de référence. Les études publiées reposent sur des contextes géographiques, des protocoles d'acquisition et des métriques d'évaluation différentes, ce qui complique la comparaison des résultats et l'interprétation des valeurs de mAP obtenues (Kunwar & Owabumoye, 2025 ; Ramos et al., 2024 ; Shobaki & Milanova, 2025).

Enfin, la détection ne constitue qu'une première étape : un modèle fiable doit également produire des mesures exploitables. Les dimensions exprimées en pixels dépendent en effet de la position et de l'orientation de la caméra, ce qui impose une étape de correction géométrique avant toute exploitation quantitative (Grimmer et al., 2025), une problématique développée en détail dans la section consacrée à l'estimation de masse.

Cette étape constitue le lien entre la détection automatique et la quantification des déchets plastiques, en permettant d'estimer des surfaces puis des masses exploitables pour le suivi environnemental.

2.3.2 Adaptation des modèles aux conditions réelles

Malgré une acquisition rigoureuse et un protocole d'entraînement robuste, un modèle performant sur un jeu de données ne conserve pas nécessairement les mêmes performances lors de son déploiement sur d'autres sites d'observation. Les variations d'éclairage, de turbidité, de saison, de débit ou encore de configuration de la caméra modifient profondément l'apparence des déchets plastiques. L'enjeu ne consiste donc plus seulement à construire un modèle performant, mais à préserver sa capacité de détection dans des environnements variés et changeants.

Repartir de zéro pour chaque nouveau site serait pourtant à la fois coûteux et peu pertinent. En effet, les premières couches d'un CNN apprennent des représentations visuelles génériques, contours, textures ou formes élémentaires, communes à la majorité des images naturelles (Jordan & Mitchell, 2015 ; LeCun et al., 2015) . Elles permettent, la généralisation du contexte paysager et l'adaptation de la détection à d'autres environnements. Ainsi le recours au *transfer learning* permet de réutiliser un modèle afin de transférer les connaissances acquises vers de nouvelles tâches (Pan & Yang, 2009). Les architectures de la famille YOLO sont généralement initialisées sur des poids génériques avant d'être spécialisées par un processus de *fine-tuning* aux caractéristiques étudié (Hussain, 2023 ; Redmon & Farhadi, 2018 ; Shukhratov et al., 2024). Cette stratégie réduit le volume de données annotées nécessaires et accélère la convergence ce qui constitue une étape incontournable pour l'adaptabilité d'un modèle à différents sites.

Le transfer learning ne garantit toutefois pas à lui seul la robustesse d'un modèle. Un réseau entraîné sur une rivière donnée reste sensible au changement de domaine (*domain shift*) lorsqu'il est confronté à des conditions d'acquisition différentes (Lavender, 2022 ; Shukhratov et al., 2024). Cette problématique est au cœur de la Domain Generalization, dont l'objectif est de développer des modèles capables de conserver leurs performances sur des environnements inédits, sans nécessiter de ré-entraînement spécifique (Zhou et al., 2022). Les différences de qualité d'eau, de végétation, de saison, de caméra ou de conditions météorologiques constituent autant de facteurs susceptibles d'altérer les performances d'un modèle entraîné dans un contexte unique. Cette difficulté est renforcée par l'hétérogénéité des jeux de données disponibles, dont les protocoles d'acquisition diffèrent fortement d'une étude à l'autre, limitant la comparaison des performances et la généralisation des modèles (Kunwar & Owabumoye, 2025 ; Ramos et al., 2024).

Au-delà de la capacité à généraliser, un système opérationnel doit également être capable d'estimer la fiabilité de ses prédictions. L'estimation de l'incertitude est aujourd'hui un domaine de recherche majeur en Deep Learning, visant à distinguer les détections fiables de celles nécessitant une interprétation prudente (Gawlikowski et al., 2022). Une prédiction réalisée dans des conditions dégradées, telles que la pluie, le brouillard, une faible luminosité ou des reflets à la surface de l'eau, ne présente pas le même niveau de confiance qu'une prédiction effectuée dans des conditions favorables (Shukhratov et al., 2024 ; Wu, 2026). L'intégration de cette information constitue un véritable outil d'aide à la décision en permettant d'identifier les situations nécessitant une validation humaine.

L'adaptation des modèles aux conditions réelles représente ainsi une étape essentielle vers le déploiement de systèmes de surveillance fiables, capables de conserver leurs performances malgré la variabilité des environnements naturels.

2.3.3 Déploiement opérationnel

Un modèle capable de détecter des déchets plastiques de manière robuste et adaptable ne constitue pas encore, à lui seul, un outil de surveillance environnementale. La détection d'objets fournit uniquement une information visuelle, alors que les gestionnaires des milieux aquatiques ont besoin d'indicateurs quantitatifs permettant de caractériser les flux de pollution. Transformer un modèle de vision par ordinateur en instrument de mesure permet ainsi de convertir les observations issues de l'image en variables physiques directement exploitables, tout en garantissant un fonctionnement autonome dans des conditions réelles de déploiement (Grimmer et al., 2025 ; Surfrider Foundation Europe, 2025).

Dans ce contexte, le recours à l'Edge AI constitue une évolution majeure. Initialement les modules installés sur le terrain ne pouvaient reposer sur des transmissions permanentes des flux vidéo vers des infrastructures Cloud en raison des contraintes de bande passante, de consommation énergétique et de latence (Shukhratov et al., 2024 ; Surfrider Foundation Europe, 2025). Pour y remédier, l'inférence est réalisée directement sur le dispositif embarqué, à l'aide d'architectures légères telles que MobileNet (Howard et al., 2017) ou des versions optimisées de YOLO (Ultralytics, 2026) . Cette mise en œuvre locale nécessite néanmoins une optimisation préalable des modèles, notamment par des techniques de compression telles que la quantification des poids et le pruning, qui réduisent l'empreinte mémoire et le coût des calculs tout en conservant des performances compatibles avec un traitement en temps réel (Gou et al., 2021 ; Howard et al., 2017). De plus, pour correspondre au problème de perte de latence lors de l'envoi, seules les images dont la détection est faite par le modèle sont envoyées vers les modules de stockage.

Ainsi le positionnement de la caméra constitue un choix méthodologique déterminant pour l'ensemble de la chaîne de traitement. Les solutions recensées dans la littérature privilégient généralement deux configurations. D'une part, les drones offrent une couverture spatiale large mais ne permettent qu'une surveillance ponctuelle et discontinue car nécessitant des recharges fréquentes (Das et al., 2026). D'autre part, les caméras fixes, généralement installées sur des ouvrages traversants tels que des ponts (Sio et al., 2022), limitent l'observation à une portion réduite du lit de la rivière de part leur positionnement à la verticale du flux. Cette seconde configuration facilite la distinction entre débris et turbidité tout en permettant le lien avec des infrastructures électriques déjà existantes. Face à ces contraintes, l'installation de caméras fixes selon un angle de vue oblique sur les berges constitue une alternative permettant d'assurer une surveillance continue sans dépendre d'un ouvrage préexistant notamment grâce à l'installation de panneaux solaires ou de batteries tout en élargissant la portion de tronçon observée par chaque dispositif.

Cette configuration oblique implique toutefois une contrepartie directe sur le plan géométrique : contrairement à une prise de vue verticale, où la distance caméra-objet varie peu, l'angle de la caméra introduit une forte variation de perspective selon la position de l'objet dans le champ. Cette contrainte renforce la nécessité d'une correction géométrique robuste pour transformer les dimensions mesurées en pixels en grandeurs physiques exploitables.

2.3.4 Estimation de masse

Comme vu précédemment, les approches traditionnelles de suivi de la pollution plastique en rivière reposaient sur des comptages manuels depuis les berges (van Emmerik & Schwarz, 2020 ; van Emmerik et al., 2018) ou par des campagnes ponctuelles d'échantillonnages par filets suivies d'analyses en laboratoire (Moore et al., 2011). Bien que ces méthodes fournissent des mesures précises de masse des débris à l'échelle de l'échantillon prélevé, elles demeurent discontinues, effectuées sur une sélection limitée d'échantillons, et limitent la compréhension des variabilités de transit des types de débris, notamment lors des épisodes de crue ou d'événements météorologiques extrêmes (Lebreton et al., 2017 ; Nakayama & Osako, 2023).

La transition vers une méthode d'estimation de la masse automatisée repose d'abord sur la conversion des dimensions observées en une surface réelle exploitable. Des travaux tels que ceux de Grimmer et al., 2025 montrent qu'à partir des boîtes englobantes produites par un modèle de convolution et d'une correction géométrique par homographie, il est possible de retrouver les dimensions 2D des objets avec une erreur résiduelle de l'ordre de deux centimètres. Une fois la surface réelle connue, l'estimation de la masse peut être formulée à partir de relations mathématiques reliant l'aire projetée, l'épaisseur supposée et la densité du matériau (Lebreton et al., 2017). Ces approches géométriques offrent un gain de place et de traitement, mais supposent la disponibilité d'informations de calibrage (hauteur et orientation de la caméra, paramètres intrinsèques de l'optique) qui constituent un prérequis parfois difficile à satisfaire pour des systèmes entièrement autonomes.

Afin de s'affranchir de cette contrainte, des solutions de réseaux de neurones pour estimer la profondeur monoculaire ont été envisagées. Dans sa revue de littérature, portant sur les systèmes à une seule caméra, **Zhang2025** soulève un défaut structurel majeur de cette famille d'approches. En effet, elles éliminent souvent toute information d'échelle dans leur processus d'estimation et perdent leur fiabilité sur des surfaces non uniformes ou réfléchissantes, ce qui les rend peu adaptées à la détection en milieu fluvial. D'autres architectures telles que SMOKE, exploitées dans les travaux de **Lu2020** pour entraîner des modèles de conduite autonome, atteignent de très bonnes performances de reconstruction 3D, mais au prix d'une complexité qui la rend incompatible avec l'Edge computing. De plus, leur entraînement reposant sur des milliers d'images annotées en trois dimensions via des maillages ou des nuages de points, il s'agit d'une ressource rare qui est inexistante dans le domaine de la détection en rivières. Ces méthodes ne permettent par ailleurs que le passage d'une représentation 2D à une représentation 3D de la forme des objets, sans apporter d'information directe sur leur masse.

Pour cela, Lusher et al., 2017, dans une étude sur la caractérisation des microplastiques ingérés par la faune aquatique, établit que les densités des polymères les plus répandus varient très peu d'un type à l'autre. Ce constat accrédite la mise en place d'une estimation de la densité moyenne des débris plastiques en fonction de leur volume apparent. Ce qui permet de compléter la chaîne de traitement en associant à chaque détection une estimation de masse à partir de la surface réelle ainsi que l'hypothèse de densité moyenne.

Ainsi, le suivi de la pollution plastique en rivière a progressivement évolué des comptages manuels et des campagnes d'échantillonnage vers des dispositifs technologiques capables d'assurer une surveillance continue. L'intégration du Deep Learning, et plus particulièrement des réseaux de neurones convolutifs, a permis de dépasser les limites des méthodes traditionnelles en automatisant la détection des déchets à partir d'images acquises in situ. Associée à des architectures embarquées adaptées aux contraintes de l'Edge AI, cette approche ouvre la voie à des systèmes de surveillance autonomes, capables de fonctionner en continu sur des sites isolés.

Toutefois, la détection seule ne suffit pas à répondre aux besoins des gestionnaires des milieux aquatiques, qui nécessitent des indicateurs quantitatifs exploitables. La conversion des dimensions mesurées en pixels vers des grandeurs physiques réelles, puis vers une estimation de masse, constitue ainsi l'étape complémentaire indispensable pour transformer un modèle de détection en véritable instrument de mesure. C'est dans cette perspective que s'inscrit notre étude, qui vise à développer et évaluer une chaîne de traitement complète, de l'acquisition des images à l'estimation de la masse des déchets plastiques, déployée directement sur des caméras installées in situ sur le bassin versant de l'Eurométropole de Strasbourg.

3 Méthodologie

3.1 Aire d'étude

Comme évoqué en introduction, les milieux aquatiques terrestres les plus exposés à la pollution plastique sont le plus souvent situés en périphérie des grandes agglomérations. Les canaux et les rivières peuvent accumuler de nombreux débris en raison des dépôts ménagers ou des zones industrielles qu'ils traversent. Il est donc pertinent d'étudier le charriage des débris plastiques au sein des voies fluviales, qui, tout au long de leur parcours, combinent des dynamiques naturelles, industrielles et urbaines, à des échelles et des densités d'habitation variables. C'est dans ce cadre que s'inscrit le choix de l'Eurométropole de Strasbourg comme zone d'étude. Cette zone frontalière avec l'Allemagne entourant la ville de Strasbourg est bordée par le Rhin, un fleuve majeur qui contribue au transport de macro-déchets de l'intérieur des terres vers les océans. Bien que de taille et de densité plus modeste que les capitales européennes souvent documentées dans la littérature, Strasbourg constitue la huitième ville de France et bénéficie d'un réseau hydrographique singulier, en effet elle est traversée par l'Ill, qui irrigue son centre-ville. La zone est aussi traversée par plusieurs de ses affluents, comme le Steingiessen, qui y prend sa source pour se jeter directement dans le Rhin, ou encore la Bruche qui se situe en amont de la ville. Ces deux rivières seront très intéressantes dans le comptage des déchets plastique car elles traversent toute deux des zones naturelles protégées. Cette configuration géographique permettra de mesurer la charge de pollution en provenance de l'amont de la ville, depuis des zones naturelles comme le Climont dans lequel prend source la Bruche, mais aussi en aval de l'Ill dans la zone protégée de la Wantzenau avec le Steingiessen.

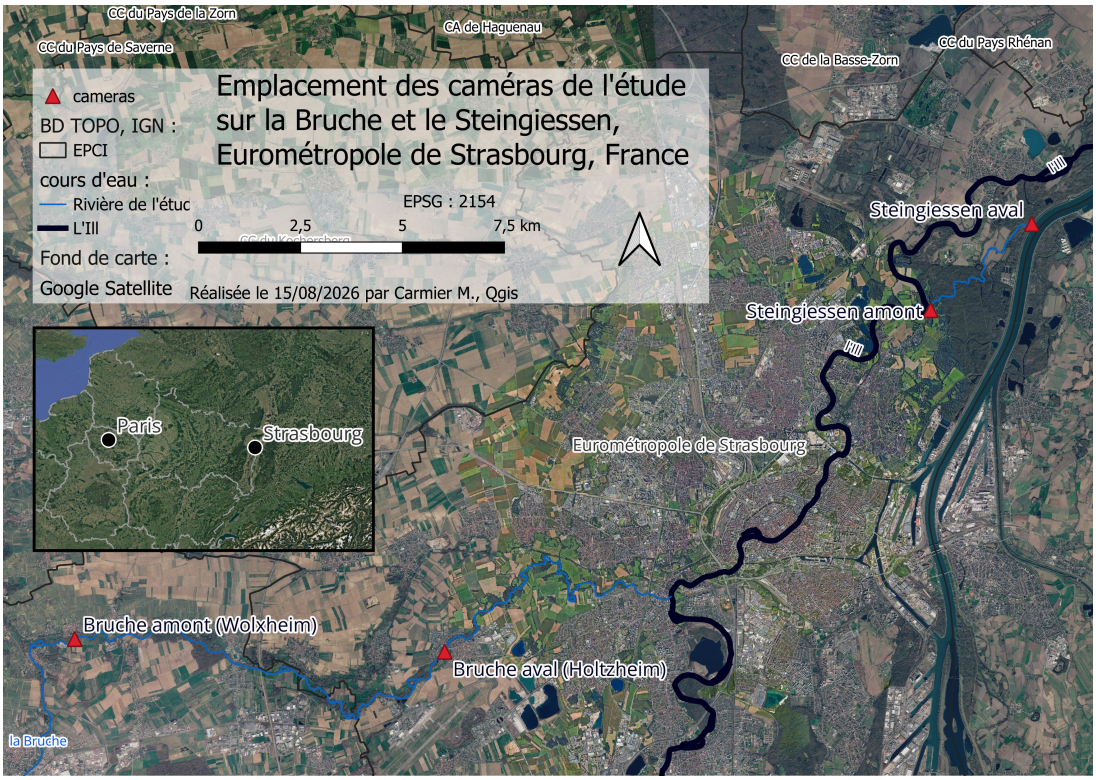


FIGURE 1 – Aire d'étude, Eurométropole de strasbourg

Notre zone d'étude se concentrera donc initialement sur ces deux cours d'eau, permettant l'évaluation des débris entrant dans l'Eurométropole avec sur la Bruche une caméra en amont de la zone, sur la commune de Wolxheim, et une vers les entrées de Strasbourg, sur la commune de Holtzheim. Deux caméras seront aussi installées en aval de l'Ill avec une caméra à l'embranchement du Steingiessen et de l'Ill et de la convergence de la rivière avec le canal du Rhin. Le projet pourra ainsi conserver en perspective l'objectif d'augmenter le nombre de caméras pour étendre le suivi automatisé à l'entièreté du bassin versant de Strasbourg.

3.2 Données

La démarche que nous avons pu extrapoler de la littérature pour l'entraînement de nos modèles repose sur l'entraînement de CNN. La première étape a donc été de constituer un jeux de données d'images annotées fiable. Lors de l'entraînement, les poids des filtres convolutifs évoluent en fonction des caractéristiques observées sur les images qui leur sont fournies successivement : cette méthode permet aux filtres d'extraire à la fois les caractéristiques du milieu observé et celles, plus spécifiques, des objets recherchés. La quantité et la qualité des annotations influencent donc directement les performances finales du modèle, sa capacité à généraliser sur différents milieux, qui est obtenue par l'extraction de caractéristiques sur un grand nombre d'images diversifiées, et sa précision, qui conditionne la limitation des faux positifs et faux négatifs biaisant les statistiques finales (Hoiem et al., 2021). Un jeu de données de taille moyenne à grande (à partir de 15 000 images) permet alors de limiter le risque que le modèle ne se fixe sur une représentation trop précise d'une rivière ou d'un débris donné. La constitution d'une base exhaustive étant fastidieuse, nous avons dans un premier temps cherché à mobiliser des données existantes issues de la littérature.

Deux grandes stratégies permettent d'adapter un modèle au contexte de terrain dans lequel il sera déployé. La première consiste à concevoir un modèle générique, capable d'apprendre des paramètres valables sur des paysages variés. Cette approche est coûteuse en données, car elle nécessite une grande quantité de données variés et un bon équilibre entre les classes annotées afin d'éviter le sur-apprentissage sur certaines catégories d'objets au détriment des autres. Elle présente néanmoins un avantage important : une meilleure compréhension intrinsèque des caractéristiques de l'image permet au modèle d'extrapoler à des configurations de milieu variées, y compris des paysages qu'il n'a jamais observés. Son principal inconvénient réside dans la puissance de calcul requise, la complexité du modèle augmentant avec la diversité des données à gérer. La seconde stratégie consiste à concevoir un modèle spécialisé, adapté à un contexte environnemental précis. Cette configuration est particulièrement utile lorsque la donnée disponible est rare : en apprenant les paramètres propres à un cadre paysager donné (luminosité, climat, faune et flore locales), le modèle se contente d'une base de données moins généraliste, donc moins volumineuse, la redondance du fond de l'image devenant secondaire au profit de la reconnaissance des cibles recherchées.

3.2.1 Données dans la littérature

La stratégie générique présente l'avantage de pouvoir s'appliquer à différents contextes, mais nécessite une quantité importante de données et de ressources de calcul. Son déploiement sur des appareils périphériques reste néanmoins possible grâce à des dispositifs permettant d'alléger le modèle et son coût d'inférence.

Nous avons recensé 18 bases de données dans la littérature pour fournir notre modèle, toutefois leur exploitation s'est heurtée à plusieurs limites majeures. *Voir table en annexe*

Les limites

Ces données ayant été principalement collectées pour l'entraînement de modèles embarqués à bord de drones, comme ceux du projet Clearbot mis en place à Hong Kong et en Inde, ou encore sur des caméras fixes positionnées à la verticale du flux par Sio et al., 2022, leurs configurations correspondent mal aux contraintes imposées par notre étude. Les caméras à orientation verticale, par exemple, font abstraction du paysage alentour en priorisant la distinction entre les débris et les turbidités ou autres mouvements de l'eau, ce qui les rend très limitantes par rapport à la structure complexe des paysages observés par nos configurations de caméra. Il en est de même pour la disposition des objets, qui sont alors uniquement vus de près et à la verticale, dans des conditions optimales, ce qui ne correspond pas à la vision oblique de nos jeux de données.

Enfin, ces bases étrangères n'ayant pas été annotées selon nos propres règles, plusieurs incompatibilités ont motivé leur mise à l'écart. La précision des boîtes englobantes en est un premier exemple : plusieurs jeux de données internationaux laissent une marge importante autour de l'objet dans leurs *bounding boxes*, alors que la typologie que nous avons adoptée exige au contraire un contour au plus près de l'objet réel. Le format d'export constitue une seconde source de divergence : ces bases utilisant des formats différents (PascalVoc, XML), une conversion vers le format YOLO a parfois été nécessaire, avec un risque d'erreur associé. Enfin, nos caméras étant positionnées sur des rivières agitées, les reflets y apparaissent plus fréquemment que sur des eaux calmes et se confondent avec les objets plastiques lorsque la résolution de caméra est insuffisante, ou selon les conditions météorologiques, un temps couvert brouillant l'image, un beau temps amplifiant les reflets. Les bases de données amplifiant cette confusion ont donc été mises de côté.

Données mobilisées

Parmi les données mobilisables, celles issues des travaux préliminaires à ce travail menés par Grimmer et al., 2025 dans le cadre de l'implantation du dispositif dans l'Eurométropole de Strasbourg se sont révélées les plus pertinentes. Cette étude a permis de constituer une base de 6 000 images annotées pour la détection de plastique en temps réel. Les résultats de Grimmer et al., 2025, corroborés par **Lin2018**, montrent par ailleurs que les images négatives et le tri temporel sur fond statique ont un impact limité, tandis que le fond de l'image peut constituer une source de sur-apprentissage.

Ces éléments confortent le choix d'un modèle spécialisé : disposer de données acquises dans l'environnement de déploiement permet au modèle de se familiariser avec les caractéristiques statiques des sites étudiés et de limiter le risque de sur-apprentissage sur les berges. Une validation sur un jeu de données provenant d'un environnement différent pourra ensuite permettre d'évaluer sa capacité de généralisation.

3.2.2 Création des nouveaux jeux de données

Face aux limites de la littérature et au manque de données mobilisables, nous avons constitué notre propre base de données. Ce jeu d'entraînement, moins complet que les bases existantes, correspond à la seconde stratégie présentée plus haut : une spécialisation de la donnée sur la zone d'étude. Entraîner le modèle sur les rivières où il sera, par la suite, déployé lui permet de se spécialiser sur le contexte propre à la zone d'étude. Le traitement des premières images, avant l'ajout de données externes, a par ailleurs permis de mieux comprendre, du point de vue opérateur, l'environnement d'installation des futures caméras et d'identifier les points de vigilance propres au contexte fluvial.

Afin de constituer des jeux de données conséquents pour l'entraînement des modèles, une partie des données ont été utilisé depuis la littérature, notamment ceux des travaux préliminaires à se sujet de Grimmer et al., 2025. Le réseau de caméras installé autour de Strasbourg ayant pu s'étendre avec l'ajout des caméras sur la Bruche que nous avons utilisées en plus de son set de données annotées qui repose sur la même méthodologie que celui développé dans cette étude. Les données collectées forment ainsi un set de données de 6000 images déjà annotées sur les caméras du Steingiessen.

La donnée à disposition pour l'entraînement des nouveaux modèles est donc plus diversifiée spatialement, avec 4 caméras au total positionnées sur le Steingiessen, ainsi que sur les deux nouveaux points de la Bruche à Holtzheim et Wolxheim. Entre les installations de ces caméras, la méthode de collecte est restée similaire avec des injections manuelles de débris devant les caméras mises en place pour la surveillance. Le passage d'éléments à détecter a été provoqué via une augmentation du nombre de débris sur le terrain en réalisant divers lancers de bouteilles en amont et récupération en aval des caméras par des passages en canoë. Cette méthode a été retenue pour obtenir des données ciblées sans avoir à traiter des heures de vidéo.

Afin de ne pas prendre en compte uniquement la différence spatiale entre les caméras, la dimension temporelle des images a elle aussi été variée entre différents horaires afin que le futur modèle soit apte à détecter des débris tout au long de l'année et quel que soit l'heure, de jour comme de nuit. Ces collectes ont été réalisées sur plusieurs dates : sur la Bruche, les acquisitions ont été réalisées de jour, le premier mars au crépuscule de 18h à 21h et le 17 avril de 14h à 18h ainsi qu'en avril pour la station de Holtzheim. Sur le Steingiessen, ces relevés ont été réalisé de jour le 4 mars 2025 et la nuit en mai 2026.



Figure 2 – Différents sites de l'étude, respectivement : Wolxheim , Holtzheim, Steingiessen amont et Steingiessen aval



Figure 3 – Sites de nuit, respectivement : Holtzheim et Steingiessen amont

TYP1 :	Wolxheim (Mars 26)	Holtzheim (Mars 26)	Holtzheim (Avril 26)	Steingiessen am (Mars 25)	Steingiessen av (Mars 25)	Somme	%
Nb Images	1605	3509	1128	4500	1514	12256	
%	13,1	28,6	9,2	36,7	12,4		
1 - cl1 : Anthropique	6709	7094	2030	8468	2405	26706	61,6
2 - cl2 : Naturel	14430	24	0	754	0	15208	35,1
3 - cl3 : Non débris	0	185	227	1010	0	1422	3,3
Somme	21139	7303	2257	10232	2405	43336	100
%	48,8	16,9	5,2	23,6	5,5		

Table 1 – Composition de la donnée de jour.

Le tableau présente la répartition des annotations diurnes collectées sur l'ensemble des sites. Les déchets anthropique (classe 1) dominent largement l'échantillon, avec 61,6 % des annotations, contre 35,1 % pour les débris naturels (classe 2) et seulement 3,3 % pour la classe 3 (kayaks, animaux), plus rare par nature sur le terrain.

TYP1 :	Holtzheim (Avril 26)	Steingiessen (Mai 26)	Somme	%
Nb Images	3831	2724	6555	
1 - cl1 : Anthropique	133	2851	2984	25,7
2 - cl2 : Naturel	6159	1817	7976	68,7
3 - cl3 : Non débris	181	468	649	5,6
Somme	6473	5136	11609	100

Table 2 – Composition de la donnée de nuit.

Répartition pour les acquisitions nocturnes. La tendance s'inverse par rapport aux données diurnes : les débris naturels (classe 2) y représentent 68,7 % des annotations, contre seulement 25,7 % pour les débris anthropique (classe 1) et 5,6 % pour la classe 3. Ceci s'explique par la réalité du terrain : sur certain point de mesure, comme c'est le cas ici, les morceaux de bois ont davantage de difficulté à s'écouler avec le flux de la rivière ce qui favorise leur accumulation et donc leur présence dans les annotations nocturnes.

Cette diversité de débris rencontrés sur le terrain, du bois flottant aux emballages plastiques en passant par les kayaks des usagers de la rivière, est illustrée en Figure 4.



Figure 4 – Classe d'objets observés (Bois,plastique, kayak) de nuit et de jour

On y observe, pour l'imagerie infrarouge et RGB, les mêmes formes de plastique et de bois, tandis que l'apparence du kayak diffère nettement d'un mode d'acquisition à l'autre. Face à cette diversité d'objets à détecter, il a été nécessaire de définir une typologie de classification structurée, permettant de distinguer ces différents types de débris selon plusieurs niveaux de granularité.

Constitution et enrichissement du jeu de données

Nous avons fixé une typologie à trois niveaux permettant de classer chaque objet détecté selon un degré de détail croissant.

- **Typologie 1 (TYP1)** : niveau le moins détaillé, qui classe les débris selon leur nature, anthropique, naturelle ou vivante. Ce premier niveau permet de détecter l'ensemble des classes présentes sur la rivière, en vue d'estimer la fréquentation par les kayakistes et les animaux, ainsi que les volumes de bois et de plastique transitant sur chaque site.

- **Typologie 2 (TYP2)** : niveau intermédiaire, qui décline les débris selon leur composition (matériau) au sein de chaque catégorie de TYP1.

- **Typologie 3 (TYP3)** : niveau le plus détaillé, qui caractérise plus finement la nature structurelle de l'objet au sein de chaque matériau, par exemple rigide ou flexible pour le plastique, ou selon les éléments constitutifs pour le bois (tronc seul, tronc et racine, tronc racine et branche) et le végétal (plante et feuille, branche).

Typo_lv1	Typo_lv2	Typo_lv3
Anthropique 1	Plastique 11	Rigide 111
		Flexible 112
	Métallique 12	
	Textile 13	
	Verre 14	
Naturel 2	Bois 21	Tronc 211
		Tronc et racine 212
		Tronc, racine et branche 213
	Végétal 22	Plante et feuille 221
		Branche 222
Non Débris 3	Kayak 31	
	Animal 32	

Table 3 – Typologie des objets à détecter. Cette table détaille la typologie réfléchie lors des travaux déjà effectués sur cette zone d'étude notamment par Grimmer et al., 2025

Règles d'annotation

Afin d'homogénéiser l'annotation et de ne pas entraîner le modèle sur des étiquettes erronées ou trop imprécises, plusieurs règles ont été établies. La première règle repose sur la visibilité de l'objet : si l'œil humain ne parvient pas à l'identifier sur une image isolée, sans le contexte de la vidéo, le modèle n'y parviendra pas non plus. Les modèles de détection ne disposent que d'un module de détection, et non d'un module de prédiction spatiale comparable à celui de l'œil humain ; un objet peut donc sembler visible dans le flux vidéo alors que sa trace est en réalité masquée par une turbidité excessive ou un reflet du soleil sur l'image fixe. Les objets non clairement visibles n'ont donc pas été annotés, afin de ne pas dégrader les performances du modèle.

La seconde règle concerne la taille des objets. Dans les versions antérieures de YOLO (de la v5 à la v8), une taille minimale d'annotation était imposée par le modèle ; ce n'est plus le cas depuis la version 26 (1er janvier 2026), que nous utilisons dans cette étude, mais les objets trop petits restent une source de confusion pour le modèle. Cette contrainte est d'autant plus importante dans notre cas que certains tests visent à rendre les modèles compatibles avec des dispositifs embarqués, et à permettre une comparaison avec les modèles plus anciens utilisés par Grimmer et al., 2025. Les objets visuellement trop petits n'ont donc pas été annotés, et un script de calcul de l'aire des boîtes englobantes a permis d'éliminer, dans les fichiers d'annotation, les lignes dont l'aire était inférieure à 10 px^2.

Outil et format d'annotation

Les premières annotations ont été réalisées avec LabelImg, un logiciel compatible avec les formats TXT, PascalVoc et XML. Les modèles de détection Ultralytics retenus pour cette étude reposent sur des boîtes englobantes (*bounding boxes*) enregistrées au format texte (.txt), où chaque ligne représente un objet selon la notation suivante :

Classe Cx Cy L H

— **Classe** : index de la classe de l'objet ;
— **Cx, Cy** : coordonnées du centre de la boîte englobante (abscisse, ordonnée) ;
— **L, H** : longueur et hauteur totales de la boîte englobante.

Le format YOLO permet en outre de faire appel à un fichier texte de référence (fichier *classes*) qui autorise le changement des intitulés de classe sans modifier les annotations déjà réalisées sur chaque image.

Annotation des images en couleurs

De plus, les objets fixes dans l'image n'ont pas été pris en compte : bien qu'ayant été labellisés sur toutes les images, les débris qui s'étaient échoués sur les berges ou embâcles naturels ont été retirés lorsqu'aucun autre débris en mouvement n'était présent sur l'image. En effet, ces débris sont à la limite entre l'objectif de détection et le fond de l'image ; s'ils sont labellisés alors que leur mouvement est nul, le modèle va apprendre par cœur leur position, ce qui va fausser les statistiques du modèle. Pour éviter que le modèle se perde en ne les indiquant pas, ils seront labellisés lorsqu'un autre objet à détecter apparaît sur l'image.

Il en est de même avec les branches, les racines et les arbres présents sur la zone de détection ; ces classes étant rares dans les jeux de données car ayant une probabilité plus faible d'apparition, et n'ayant pas pu être artificiellement augmentées comme avec les plastiques, enlever leur rare apparition aurait empêché le modèle de les détecter correctement. Il en résulte que la principale source d'apparition de branches et de végétation en tout genre est les berges, ce que le modèle doit absolument éviter d'apprendre. En effet, si le modèle apprend à associer des items sur les berges avec ceux défilant dans les rivières, il ne parviendra pas à faire la distinction entre les deux et ne pourra alors pas détecter les débris qui se trouveront proches des berges ou qui s'y échoueront.

Annotation des images infrarouges

Lors de la labellisation des images nocturnes, nous avons conservé la même typologie et les mêmes règles. Toutefois, une attention plus accrue a dû être portée aux visuels des objets, car, de nuit, des insectes passant près de la caméra et formant des silhouettes floues peuvent être confondus avec des branches. Ces faux positifs, plus souvent observés que les branches et les plastiques sur le terrain, pourront potentiellement confondre le modèle en situation réelle, quitte à détériorer les résultats sur un modèle mixte.

3.2.3 Choix de la résolution des images

Les données utilisées pour l'entraînement du modèle ont été récoltées à partir de différents appareils d'acquisition, bien qu'elles aient été installées dans un même secteur d'étude chaque site possède différents modèles de caméra. De ce fait, les différents jeux de données annotés ne possèdent pas tous la même résolution d'acquisition, la standardisation des résolutions constitue donc une étape nécessaire de la préparation des données avant l'entraînement du modèle. Elle permet de disposer d'entrées homogènes pour le réseau y compris pour les futures caméras qui viendront s'ajouter au dispositif, tout en établissant un compromis entre la quantité d'information conservée dans l'image et le coût computationnel associé à son traitement.

Dans le cas présent, l'utilisation de données provenant de quatre caméras réparties sur deux sites limite la variabilité des résolutions : deux configurations de caméra seulement ont été utilisées, correspondant aux résolutions présentées dans le tableau 5 (1).

La réduction de la résolution des images crée pourtant quelque contrainte, car les objets recherchés sont des déchets flottants dont la taille est variable. Une réduction trop importante de la résolution risque donc de diminuer le nombre de pixels représentant un objet et d'en dégrader la détectabilité. Au contraire, conserver les résolutions originales augmente le nombre de pixels à traiter, et donc le coût computationnel du modèle, pour des surfaces d'image qui ne correspondent pas nécessairement à des éléments réellement observables et qui peuvent confondre le modèle comme des feuilles dont la réflectance à la surface de l'eau permet une confusion avec les items plastiques peu apparents. Le choix de la résolution a donc été réalisé à partir de la taille minimale des objets annotés visibles dans les images, l'objectif est de déterminer jusqu'à quel niveau les images peuvent être réduites pour que les plus petits objets soient toujours apparents.

Détermination d'un seuil minimal de représentation des objets

Une première étape consiste à déterminer la taille minimale en pixels à partir de laquelle un objet peut être considéré comme suffisamment apparent pour être détecté par le modèle. En effet, des objets trop petits peuvent soit se perdre et limiter le nombre de détections vraies en sous-estimant leur passage, soit, en se confondant avec des feuilles ou des insectes, augmenter le nombre de faux positifs. Il a donc fallu, par l'étude des plus petits objets observables dans nos annotations manuelle et de leur distribution, déterminer un seuil valable à partir duquel une détection pouvait être considérée comme fiable.

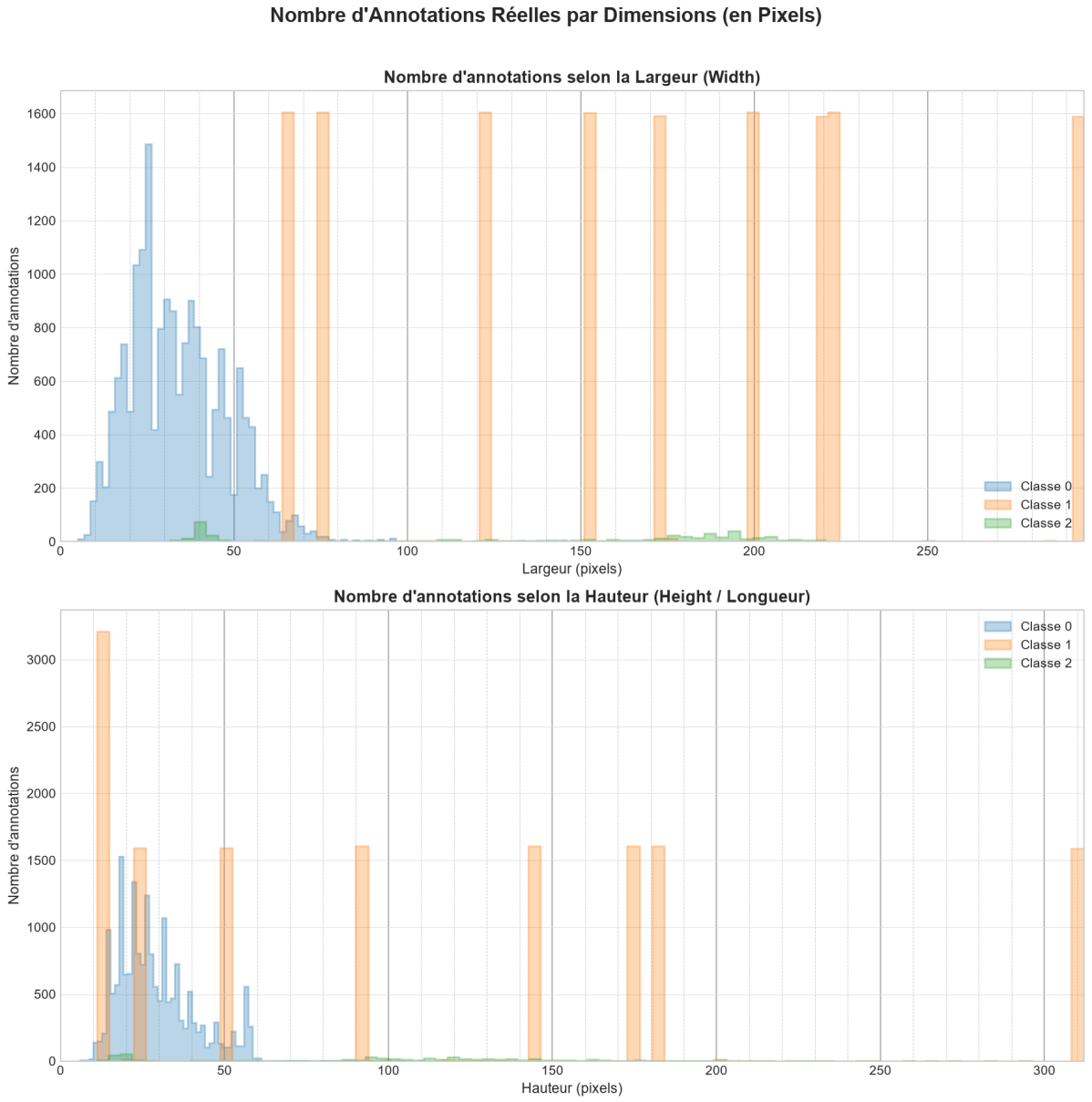


FIGURE 5 – distribution des dimensions des annotations

Dans la figure 5 la distribution des dimensions des boîtes englobantes permet d'observer que de nombreuses annotations, notamment dans la classe 1, présentent des dimensions très petites avec une part non negligeable de données inférieure à 20 pixels et plusieurs pic de données la plus grande concentration se situant autour de 35 pixels.

Le plus petit objet retenu comme valide, en supposant que les objets encore plus petits correspondent à des erreurs d'annotation ou à des confusions trop évidentes pour être exploitables par le modèle, présente une taille d'environ 18×18 pixels, ce qui correspond au premier pic de largeur et de hauteur observé sur la distribution. Cette valeur a été retenue pour une comparaison avec le premier quartile de la distribution, situé autour de 26×21 pixels. Ces deux valeurs permettant d'étudier deux scénarios distincts. Le premier scenario ayant pour but de conserver le plus grand nombre d'objets possible, en ne supprimant que les valeurs qui ne sont pas visuellement significative, elle conduit à un facteur de réduction faible qui limite la réduction de la résolution de l'image. Au contraire le second scénario s'appuie sur les premiers quartiles des dimensions des objets. Cette valeur permet de caractériser une taille plus représentative de la majorité des petits objets, sans être directement déterminée par les annotations extrêmes les plus petite. Elle permet d'offrir une comparaison et d'évaluer le nombre d'objets qui seraient perdus si la résolution était réduite de manière à conserver au minimum cette taille. Cette distinction est importante, car une réduction calculée uniquement à partir du plus petit objet peut conduire à une résolution finale inutilement élevée. À l'inverse, choisir une réduction trop importante pourrait supprimer une proportion significative des objets de petite taille et modifier la distribution des données utilisées pour l'apprentissage. Il est également nécessaire de mentionner l'effet de cette approche par projection : le calcul du quartile ayant été réalisé séparément sur les distributions de largeur et de hauteur, il aboutit à deux valeurs distinctes (26 et 21 pixels), alors que le choix de 18 pixels résulte d'une observation empirique d'un pic commun aux deux distributions. Le seuil du quartile est donc potentiellement mieux adapté à chaque axe pris séparément que la valeur unique de 18 px. Le tableau 4 présente la proportion d'annotations qui ne seraient plus conservées selon différents modes de suppression, selon qu'une seule dimension (L ou H) se situe sous le seuil déterminé.

Q_plastique <	Q_tot_plastique	%plastique <	%dimension	Mode
Annotations <1er Quartile				
7560	30564	24,7	13,0	L
7560	30564	24,7	13,0	H
Annotations <18 px				
2478	30564	8,1	4,2	L
3969	30564	13,0	6,8	H

Table 4 – Éléments les plus petits

Les résultats montrent que le choix d'un seuil correspondant au premier quartile entraîne une suppression plus importante d'annotations que le seuil de 18 pixels. Néanmoins, cette perte reste à mettre en regard du gain obtenu sur la réduction de la résolution. À l'inverse, le seuil de 18 pixels permet de conserver davantage d'annotations, mais impose une réduction plus limitée des images.

L'analyse permet ainsi de définir une contrainte de résolution : après réduction, les plus petits objets conservés doivent disposer d'une représentation suffisante pour rester détectables par le modèle. La comparaison des dimensions retenues par les deux méthodes permet d'observer l'écart entre le seuil empirique fixe et le seuil basé sur le premier quartile, détaillé dans le tableau 5 (2).

Calcul du facteur de réduction.

Grâce à nos deux méthodes, les tailles minimales de nos objets dans les jeux de données ont pu être estimées ; pour calculer le facteur de réduction de la résolution des images, il restait cependant à déterminer celle des items après réduction. Plusieurs travaux comparant les performances de détection de différentes versions de YOLO sur des petits objets retiennent un seuil relatif à la taille de l'image, de l'ordre de 1 % de sa dimension (Tariq & Javed, 2025). Appliqué à nos résolutions natives, celles sur lesquelles les annotations ont été réalisées, ce seuil se calcule sur la dimension la plus contraignante de chaque caméra :

$$0{,}01 \times 1296 \text{ px} \approx 13 \text{ px} \quad \text{(Bruche) et} \quad 0{,}01 \times 1080 \text{ px} \approx 11 \text{ px} \quad \text{(Steingiessen)}$$

Entre ces deux valeurs, la plus restrictive, celle du Steingiessen, la caméra à la résolution la plus faible, a été retenue afin que le seuil reste valable sur l'ensemble du dispositif, puis arrondie à la valeur ronde inférieure la plus proche pour simplifier les calculs et harmoniser le traitement entre les deux caméras, soit 10 px. Ce choix reste des lors plus strict que celui suggéré par la littérature : il vise avant tout à écarter les détections aberrantes plutôt qu'à maximiser le rappel sur les très petits objets. Ce repère issu de la littérature ne concerne toutefois que des versions antérieures de YOLO, la version 26 étant trop récente (janvier 2026) pour avoir fait l'objet d'une telle évaluation ; il ne constitue donc qu'une estimation de cohérence, et non une valeur de référence directement applicable à notre pipeline. Ce choix est par ailleurs cohérent avec le mécanisme STAL (*Small-Target-Aware Label Assignment*) intégré à YOLO26 et présenté dans l'état de l'art (**Jocher2026**), qui garantit qu'un objet de petite taille reçoive toujours au moins une assignation positive lors de l'entraînement, même lorsque sa taille est inférieure au pas minimal du réseau. Ce mécanisme renforce la pertinence du choix de cette architecture pour cette étude, en limitant le risque que les plus petits objets conservés après réduction de résolution soient ignorés par le processus d'apprentissage. L'importance des objets trop petits étant par ailleurs moindre par rapport à la priorité d'une détection vraie, ce seuil constitue un compromis entre la conservation des très petits objets et des débris d'une part, et la réduction de la résolution nécessaire à l'allègement du traitement d'autre part.

À partir de la taille minimale retenue, un facteur de réduction est ensuite calculé pour chacune des dimensions de l'objet. Pour une dimension donnée, le facteur de réduction est défini comme :

$$F_x = \frac{R_x}{R_{x,\text{cible}}} \text{ et } F_y = \frac{R_y}{R_{y,\text{cible}}},$$

où R_x et R_y correspondent respectivement à la largeur et à la hauteur initiales de l'objet dans l'image, tandis que $R_{x,\text{cible}}$ et $R_{y,\text{cible}}$ représentent les dimensions minimales souhaitées après réduction.

La réduction étant appliquée simultanément selon les deux axes, le facteur effectivement retenu doit respecter la contrainte la plus restrictive. Le facteur de réduction final est donc déterminé par le minimum des deux valeurs

$$F = \min(F_x, F_y).$$

Cette approche permet de garantir que la dimension la moins favorable de l'objet reste au-dessus du seuil fixé après réduction. Autrement dit, la réduction n'est pas déterminée indépendamment sur les axes horizontal et vertical : elle est contrainte par l'axe conduisant à la plus faible capacité de réduction. Dans le cas du seuil de 10 pixels, l'objet de dimension 18×18 pixels conduit à des facteurs de réduction de $1,8$ selon les deux dimensions. Le facteur retenu est donc $1,8$. Pour l'objet correspondant au premier quartile, de dimension 26×21 pixels, les facteurs obtenus sont respectivement de $2,6$ et $2,1$. Le facteur limitant est alors $2,1$, comme détaillé dans le tableau 5 (3). Ces valeurs permettent de définir deux niveaux de réduction possibles. Une réduction d'un facteur de $1,8$ permet de préserver la représentation des objets les plus petits observés dans les annotations, tandis qu'une réduction d'environ $2,1$ est compatible avec une stratégie fondée sur le premier quartile.

Le choix final doit cependant également tenir compte de la résolution effectivement disponible après réduction et de la nécessité de disposer d'une valeur commune pour les différentes caméras.

Détermination de la résolution finale.

Le facteur de réduction calculé précédemment est ensuite appliqué aux résolutions natives des deux types de caméras. Cependant, la valeur théorique obtenue ne constitue pas directement la résolution finale utilisée pour le traitement. Une résolution adaptée au pipeline doit également faciliter le traitement informatique et conserver une certaine homogénéité entre les différentes configurations. La résolution finale a donc été déterminée en appliquant le facteur de réduction retenu, puis en sélectionnant des dimensions proches de ces valeurs et adaptées au traitement par le réseau. Les dimensions ont notamment été choisies de manière à rester compatibles avec une réduction sur la puissances de deux la plus proche, ce qui facilite les opérations successives de sous-échantillonnage réalisées dans les architectures convolutives. Les résolutions retenues sont présentées dans le tableau 5 (4). Cette étape permet ainsi de réduire significativement la quantité de données traitées tout en conservant une représentation suffisante des déchets de petite taille. Le choix de la résolution permet un compromis entre la préservation de l'information utile à la détection et la réduction du coût de calcul.

Résolutions des caméras : (1)

Bruche	2304	1296
Steingiessen	1920	1080

Résolution de l'objet le plus petit : (2)

fixe :	L	H	Quartile :	L	H
	18	18	Cl0	26	21
			Cl1	66	19
			Cl2	178	92

Calcul du facteur de réduction : (3)

Résolution plus petit objet :	10	10	10	10
Facteur de réduction	1,8	1,8	2,6	2,1
Plus petit	1,8		2,1	

Calcul de la résolution : (4)

	width	height	width	height
Bruche	1280,0	720,0	1097,1	617,1
Steingiessen	1066,7	600,0	914,3	514,3

Multiple de 32 :

Bruche	1280	704	1088	608
Steingiessen	1056	576	896	512

TABLE 5 – Calibration des résolutions de caméras

3.3 Entraînement des modèles

La détermination de la résolution optimale des images constitue une étape préalable indispensable. Elle permet de fixer le format d'entrée du réseau et conditionne directement le compromis entre précision de détection et charge de calcul supportable par le système embarqué visé. C'est dans cette perspective que le choix s'est porté sur l'architecture YOLO26, dont les avantages pour l'application visée, orientation vers l'embarqué, suppression du NMS, nouvelles têtes de détection adaptées aux petits objets, ont déjà été détaillés dans l'état de l'art.

Métriques d'évaluation

L'évaluation des performances de détection repose sur un ensemble de métriques standard en vision par ordinateur, dont le calcul s'appuie sur la notion d'intersection sur union (*Intersection over Union*, IoU). Pour une boîte englobante prédite et une boîte annotée de référence, l'IoU mesure leur degré de recouvrement spatial comme le rapport entre l'aire de leur intersection et l'aire de leur union :

$$\text{IoU} = \frac{\text{Aire}(B_{\text{préd}} \cap B_{\text{vérité}})}{\text{Aire}(B_{\text{préd}} \cup B_{\text{vérité}})}$$

Une détection n'est considérée comme correcte (vrai positif) que si son IoU avec une boîte de référence dépasse un seuil fixé au préalable ; à défaut, elle est comptée comme un faux positif, et toute boîte de référence non appariée à une détection constitue un faux négatif.

À partir de cette classification, la précision P et le rappel R sont définis respectivement comme la proportion de détections effectivement correctes parmi l'ensemble des détections produites, et la proportion d'objets réellement présents qui ont été retrouvés par le modèle :

$$P = \frac{VP}{VP + FP}, \qquad R = \frac{VP}{VP + FN}$$

où VP, FP et FN désignent respectivement les vrais positifs, les faux positifs et les faux négatifs.

Ces deux métriques évoluent souvent en sens opposé selon le seuil de confiance retenu pour valider une détection : un seuil bas augmente le rappel au prix de la précision, et inversement. Le score $F1$ synthétise ce compromis en calculant leur moyenne harmonique :

$$F1 = 2 \cdot \frac{P \cdot R}{P + R}$$

La précision moyenne (*Average Precision*, AP) résume, pour une classe donnée, l'aire sous la courbe précision-rappel obtenue en faisant varier le seuil de confiance des détections. Le *mean Average Precision* (mAP) en constitue la moyenne sur l'ensemble des classes. Le mAP50 correspond à cette moyenne calculée pour un seuil d'IoU fixé à 0,5 : une détection est jugée correcte dès lors qu'elle recouvre au moins la moitié de la boîte de référence, ce qui en fait une métrique relativement tolérante à l'imprécision de localisation. Le mAP50-95, plus exigeant, correspond à la moyenne du mAP calculé sur une série de seuils d'IoU allant de 0,5 à 0,95 par pas de 0,05 : il pénalise davantage les boîtes englobantes mal ajustées et constitue à ce titre un indicateur plus strict de la qualité de localisation du modèle, complémentaire du mAP50 qui reflète surtout sa capacité de détection globale.

3.3.1 Méthode Statistique

Au-delà du choix des hyperparamètres, la fiabilité des résultats obtenus dépend directement du protocole expérimental mis en œuvre pour entraîner et évaluer le modèle. Ces expérimentations ont reposé sur deux architectures distinctes selon l'objectif poursuivi pour un entraînement donné. Pour la validation croisée par K-fold, une architecture YOLO26, vue comme la meilleure d'après l'état de l'art, est retenue en configuration de production. En complément, afin d'évaluer l'apport de l'enrichissement du jeu de données par nos propres données annotées, et afin de réaliser une comparaison avec les résultats obtenus par Grimmer et al., 2025, plusieurs modèles YOLOv8n ont été entraînés. Ces derniers ont ainsi pu servir de comparatif entre différentes stratégies d'initialisation des poids (poids déjà entraînés, initialisation aléatoire, poids pré-entraînés sur COCO), l'architecture YOLOv8n étant celle utilisée par Grimmer et al., 2025 dans son étude de référence.

La stratégie d'initialisation des poids du réseau constitue à ce titre l'un des facteurs déterminants à prendre en compte : plusieurs approches ont donc été confrontées afin d'évaluer laquelle permet la meilleure adaptation du modèle à notre contexte d'étude, entre un apprentissage entièrement nouveau sur nos propres données, une reprise des poids déjà spécialisés du modèle de référence de la littérature, et une reproduction fidèle du protocole d'origine servant de point de comparaison direct.

Stratégies d'initialisation

- **Configuration 1** : initialisation aléatoire (`yolov8n.yaml`), sans recours à des poids pré-entraînés, le modèle est entraîné intégralement sur notre jeu de données consolidé (D2), afin de permettre aux filtres convolutifs de se spécialiser uniquement sur les caractéristiques visuelles propres à notre contexte fluvial ;

- **Configuration 2** : fine-tuning, c'est-à-dire un ré-entraînement à partir des poids déjà spécialisés du modèle de Grimmer et al., 2025 (`last.pt`, issu de son entraînement *TRAIN3*), ajustés sur notre jeu de données nouvellement labellisé (D2) ;

- **Configuration 3, reproduction de la baseline de la littérature** : reproduction du protocole d'entraînement de Grimmer et al., 2025, à savoir un fine-tuning à partir des poids génériques pré-entraînés sur COCO (`yolov8n.pt`), mené sur son jeu de données d'origine (D1). Une variante avec suppression de la classe la plus sujette à confusion a également été testée.

Pour les configurations 2 et 3, aucun gel de couches (*layer freezing*) n'a été appliqué : l'intégralité du réseau reste entraînable durant l'ajustement, seul le taux d'apprentissage est réduit (lr0 = 0,0005 contre 0,01 pour un entraînement complet), conformément à la pratique usuelle visant à limiter l'oubli catastrophique (*catastrophic forgetting*) des représentations déjà apprises par le modèle source.

Organisation des jeux de données

Afin d'évaluer l'impact du partitionnement spatial et temporel sur les performances du modèle, deux stratégies de répartition du corpus d'images annotées sont comparées. Le premier jeu (D1) reprend le découpage défini par Grimmer et al., 2025, retenu pour la Configuration 3 : le site de Wolxheim sert à la validation, Holtzheim (avril) au test, tandis que l'entraînement s'appuie sur Holtzheim (mars) et Steingiessen (2025). Le second jeu (D2), associé aux Configurations 1 et 2, adopte notre protocole : un ensemble regroupant Holtzheim (mars), Steingiessen (2025) et Wolxheim (mars) est subdivisé aléatoirement en entraînement et validation (80/20), le test étant confié aux sites de Holtzheim (avril) et Steingiessen (aval). La répartition détaillée de ces sous-ensembles est présentée dans le tableau 6.

Split	D1 — composition	%img	%item	D2 composition	%img	%item
Train	Holtzheim (mars) & Steingiessen (2025)	77,7	46,0	Tirage aléatoire (pool D1 train+val)	64,6	n.c.
Val	Wolxheim (mars)	13,1	48,8	Tirage aléatoire (pool D1 train+val)	16,1	n.c.
Test	Holtzheim (avril)	9,2	5,2	Holtzheim (avril) + Steingiessen (aval)	21,6	10,8

Table 6 – Organisation comparée des jeux de données D1 (split par site, Grimmer et al.) et D2 (split aléatoire 80/20, notre étude). n.c. : proportion d'items non communiquée pour ce split.

Pour la Configuration 2, les poids initiaux (*last.pt*) proviennent d'un modèle entraîné par Grimmer et al., 2025 sur des données incluant le site de Steingiessen, qui constitue également l'un des trois sites composant le pool d'entraînement/validation de D2.

Le split aléatoire retenu pour D2 consiste à répartir l'ensemble des images disponibles entre les jeux d'entraînement et de validation par tirage aléatoire selon une répartition en 80/20, indépendamment de leur origine (site, date d'acquisition, conditions de prise de vue), contrairement au split organisé par acquisition de D1, où l'appartenance d'une image à l'entraînement ou à la validation est déterminée à l'avance selon le lieu où elle a été acquise. Cette approche présente un intérêt méthodologique à plusieurs égards. Premièrement, en mélangeant l'ensemble des sites disponibles (Holtzheim, Steingiessen, Wolxheim) au sein d'un même pool avant tirage, elle permet de maximiser la diversité des conditions d'acquisition représentées à la fois dans le jeu d'entraînement et dans le jeu de validation, ce qui est particulièrement utile lorsque le nombre de sites disponibles reste limité et que chacun présente des caractéristiques propres (turbidité, luminosité, configuration de berges).

Deuxièmement, cette méthode permet d'exploiter la totalité des annotations pour l'apprentissage du modèle, sans réserver un site entier à la seule validation comme c'est le cas dans une organisation par site. Troisièmement, le split aléatoire constitue un complément méthodologique à l'organisation par site : les deux approches, appliquées respectivement à D2 et D1, permettent d'examiner la performance du modèle sous deux régimes de validation différents, l'un favorisant la diversité interne des sous-ensembles, l'autre la capacité de généralisation à un site totalement absent de l'entraînement.

3.3.2 Méthode du K-fold

Afin d'évaluer la robustesse d'un modèle, il est nécessaire d'estimer ses performances sur des données qui n'ont pas été utilisées lors de son entraînement. Le jeu de test permet notamment d'évaluer les performances finales du modèle sur des données indépendantes de l'apprentissage. Cependant, une évaluation réalisée sur une seule répartition des données peut dépendre des échantillons sélectionnés et ne pas refléter suffisamment la robustesse du modèle. La validation croisée par *K-fold* constitue une méthode couramment utilisée pour limiter cette dépendance.

Le principe consiste à diviser le jeu de données en K sous-ensembles, appelés *folds*. À chaque itération, l'un des sous-ensembles est utilisé pour la validation tandis que les autres sont utilisés pour l'entraînement. Le processus est répété jusqu'à ce que chaque sous-ensemble ait été utilisé comme jeu de validation. Les performances obtenues lors des différentes itérations peuvent alors être comparées et agrégées afin d'obtenir une estimation plus robuste des performances du modèle. Cette méthode est particulièrement intéressante lorsque la quantité de données disponible est limitée, puisqu'elle permet d'utiliser une grande partie des données pour l'entraînement tout en faisant varier les données utilisées pour l'évaluation.

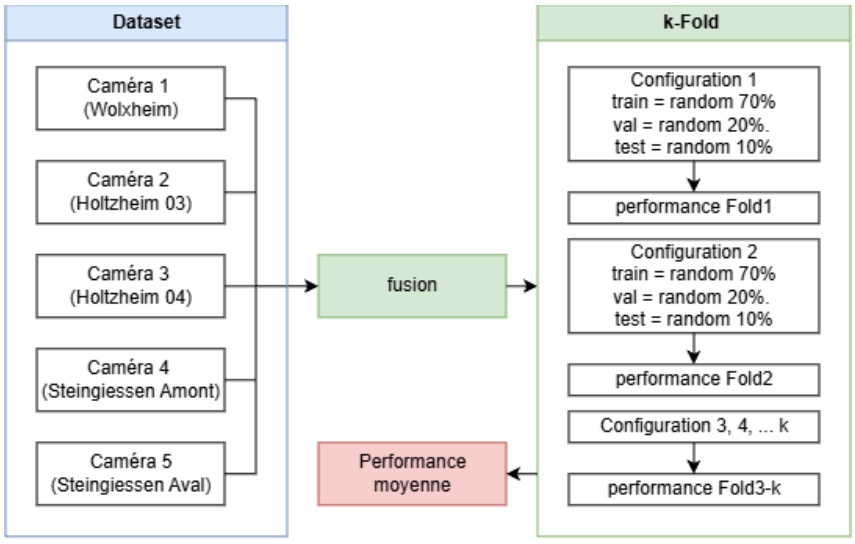


FIGURE 6 – Principe général de la validation croisée par K-fold.

Dans sa forme classique, la validation croisée par K-fold repose sur une répartition aléatoire des données entre les différents sous-ensembles. Cette approche suppose que les observations sont suffisamment indépendantes et que chaque sous-ensemble reste représentatif de la distribution des données étudiées. Or, cette hypothèse n'est pas pleinement vérifiée dans le cadre de cette étude.

Les images utilisées proviennent de plusieurs caméras fixes, chacune étant associée à un site et à des conditions d'acquisition qui lui sont propres. Les images d'une même caméra partagent ainsi des caractéristiques communes liées notamment au paysage observé, à la géométrie de prise de vue, aux conditions d'éclairement et aux conditions météorologiques. De plus, plusieurs images peuvent provenir d'une même séquence vidéo : deux images successives peuvent donc présenter des contenus très similaires tout en étant considérées comme des observations distinctes.

Une répartition aléatoire de l'ensemble des images pourrait alors conduire à placer dans des jeux différents des images provenant d'une même séquence, d'une même date d'acquisition ou d'une même caméra. Le modèle pourrait de fait être évalué sur des images très proches de celles utilisées lors de son entraînement. De bonnes performances pourraient alors refléter en partie la similarité entre les données d'entraînement et de validation, plutôt que la capacité du modèle à détecter de nouveaux objets ou à généraliser à des conditions d'observation différentes.

Cette distinction est particulièrement importante dans le contexte de cette étude. L'objectif n'est pas uniquement d'obtenir de bonnes performances dans les conditions d'une caméra donnée, mais également d'évaluer la robustesse du modèle face à des conditions d'acquisition différentes. Les différences entre les caméras constituent donc une source de variabilité qu'il est nécessaire de préserver plutôt que de diluer dans un mélange aléatoire. Les différents sites confrontent notamment le modèle à des paysages, des géométries de prise de vue, des niveaux d'ensoleillement et des conditions météorologiques différents. Même si la généralisation à une nouvelle caméra constitue un objectif secondaire par rapport à la capacité générale de détection, elle apporte une information supplémentaire sur la robustesse et l'adaptabilité du modèle.

Il est donc nécessaire de conserver la distinction entre les différents ensembles de données lors de l'évaluation, plutôt que de considérer l'ensemble des images comme un jeu de données homogène. Une adaptation du principe de validation croisée est alors mise en place afin de conserver des ensembles d'images provenant de sites distincts et d'évaluer le modèle sur des conditions d'acquisition qui n'ont pas été utilisées lors de son entraînement. Cette organisation permet de mieux caractériser sa capacité de généralisation spatiale tout en limitant les risques de fuite d'information entre les jeux d'entraînement et d'évaluation. La méthode retenue dans le cadre de cette étude est présentée dans la section suivante.

Adaptation de la validation croisée

Plutôt que de répartir les images aléatoirement entre les différents folds, cette méthode va les préserver en conservant les stack d'images provenant d'une même caméra. Cette répartition permettra de maintenir les caractéristiques propres à chaque paysage et d'évaluer le modèle sur des conditions d'acquisition différentes de celles rencontrées lors de son entraînement. La validation croisée est alors réalisée manuellement en faisant varier la répartition des jeux de données : chaque configuration est constituée de trois jeux d'entraînement, d'un jeu de validation et d'un jeu de test, ce qui permet d'étudier l'influence de la composition des jeux de données sur l'apprentissage. En effet, les performances d'un modèle ne dépendent pas uniquement de l'architecture utilisée ou de ses hyperparamètres, mais également des données sur lesquelles il est entraîné : une configuration peut favoriser l'apprentissage de certaines caractéristiques, tandis qu'une autre fournit des informations moins représentatives ou plus difficiles à exploiter. Les 20 combinaisons possibles ont été testées (tableau 7).

Cette adaptation présente néanmoins une différence importante avec une validation croisée classique. Dans un K-fold traditionnel, les différents folds sont des sous-échantillons aléatoires d'une même distribution globale de données : ils sont donc statistiquement comparables entre eux, ce qui permet d'agréger les métriques de performance obtenues sur chaque fold en une estimation unique et robuste des performances du modèle. Ici, au contraire, les modèles sont entraînés sur des combinaisons différentes de paysages, correspondant à des distributions de la données distinctes plutôt qu'à des échantillons interchangeables. Les performances obtenues sur chaque configuration ne peuvent donc pas être agrégées de la même manière que dans un K-fold classique, puisqu'un score obtenu sur une configuration donnée ne reflète pas la même réalité qu'un score obtenu sur une autre : chaque configuration teste une capacité de généralisation différente, propre à la combinaison de paysages étudier.

Par ailleurs, un moyennage des poids des différents modèles entre eux ne serait de toute façon pas envisageable, indépendamment de cette question d'agrégation des métriques : les filtres convolutionnels s'ajustant en fonction des images présentées au réseau, les poids obtenus à l'issue d'entraînements réalisés sur des ensembles différents peuvent correspondre à des configurations distinctes, sans qu'il soit possible de déterminer si les filtres correspondants représentent les mêmes caractéristiques ou des états équivalents de convergence.

Cette méthode ne permet donc pas de produire un modèle unique de synthèse, mais elle permet d'évaluer l'influence des différentes configurations de données sur les performances du modèle. La conservation d'un jeu de test correspondant à un paysage non utilisé lors de l'entraînement permet également de mesurer la capacité du modèle à généraliser à une condition d'acquisition différente, les résultats obtenus sur les différentes configurations pouvant être comparés afin d'identifier l'influence de la composition des jeux d'entraînement, de validation et de test.

ID	Site validation	Site test	ID	Site validation	Site test			
11	W3	H3	31	H4	Sam	51	Sav	W3
12	W3	H4	32	H4	Sav	52	Sav	H3
13	W3	Sam	33	H4	W3	53	Sav	H4
14	W3	Sav	34	H4	H3	54	Sav	Sam
21	H3	H4	41	Sam	Sav			
22	H3	Sam	42	Sam	W3			
23	H3	Sav	43	Sam	H3			
24	H3	W3	44	Sam	H4			

Table 7 – Configurations de la validation croisée par site : pour chaque combinaison, l'entraînement utilise les trois sites restants. Sam : Steingiessen amont ; Sav : Steingiessen aval ; H3 : Holtzheim mars ; H4 : Holtzheim avril ; W3 : Wolxheim mars.

Cette première expérimentation met ainsi en évidence l'importance de la stratégie de séparation des données dans l'entraînement d'un modèle de détection, et la nécessité de considérer la dimension temporelle des acquisitions, puisque les images issues d'une même caméra peuvent présenter des dépendances liées à leur période d'acquisition. Une seconde stratégie de séparation des données est donc étudiée afin de mieux prendre en compte cette dimension temporelle et de limiter les similitudes entre les images utilisées pour l'apprentissage et celles utilisées pour l'évaluation.

3.3.3 Méthode temporelle

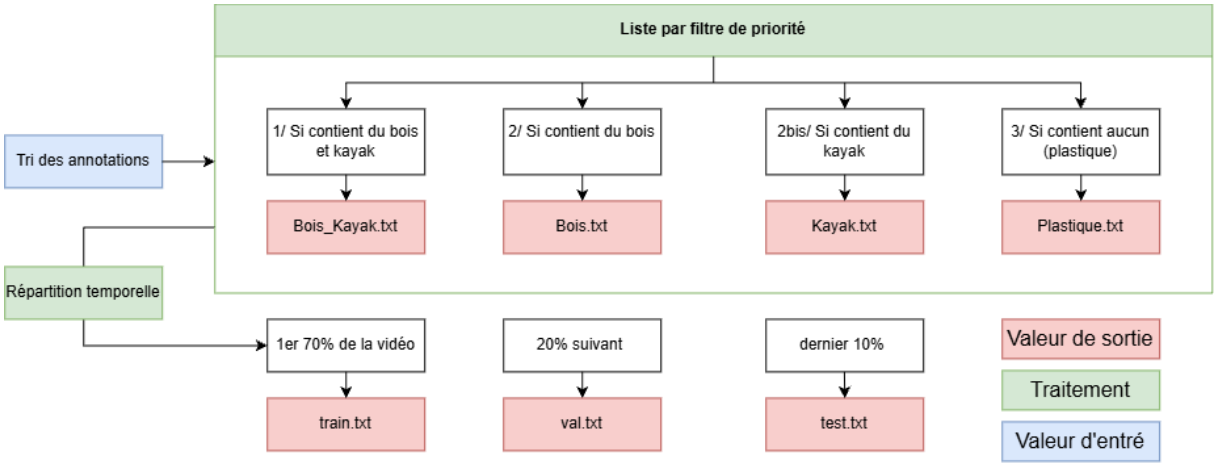


FIGURE 7 – méthode temporelle

Dans le cadre de la modélisation de données dépendantes du temps, comme le sont nos vidéos, le découpage aléatoire traditionnel (*random split*), présenté plus tôt, présente une limite méthodologique : il induit une fuite d'informations (*data leakage*) en mélangeant indistinctement des observations issues de différentes périodes chronologiques. Le modèle risque alors, s'il observe une image issue d'une séquence sur laquelle il a été entraîné, d'utiliser des observations proches et quasi identiques pour prédire des événements déjà vus. Cette interconnexion entre les jeux d'apprentissage et de test conduit à une surestimation des performances de l'algorithme et masque son incapacité réelle à reconnaître un objet qu'il n'a pas déjà vu.

Pour garantir la validité et la fiabilité de nos résultats, la méthode du split temporel, ou découpage chronologique, est appliquée lors de la création des jeux de données. Cette approche consiste à établir une frontière temporelle entre les sous-ensembles d'entraînement, de validation et de test : le modèle est entraîné sur le début de la vidéo, tandis que le test s'effectue sur la fin, garantissant une séparation suffisante entre les images pour évaluer le modèle sur des configurations inédites.

Toutefois, nos jeux de données ne contenant pas la même quantité d'images pour chaque classe, et leur distribution dans le temps n'étant pas homogène, une simple division en split 70/20/10 dégraderait les paramètres d'entraînement et d'évaluation pour les classes les moins bien représentées. Selon le constat effectué lors de l'analyse de la distribution des annotations par caméra, nous avons pu observer que les annotations de débris plastique concernent 61,6 % des annotations effectuées, contre seulement 35,1 % pour les débris naturels et 3,3 % pour le kayak. La figure 8 présente la répartition des annotations dans le temps par vidéo. On y observe que la classe kayak, présente uniquement dans les vidéos du Steingiessen amont et dans celles de la Bruche, risquerait de ne pas être représentée dans le jeu de test avec un découpage aléatoire classique, et qu'une seule image représenterait alors la catégorie bois.

Pour pallier ce problème, nous distinguons chaque classe les unes des autres en séparant les annotations de bois, de kayak et de plastique. Comme il est possible que deux classes apparaissent simultanément sur une même image, nous priorisons à chaque fois la classe la moins bien représentée : si une image est annotée avec du plastique et du bois, elle est catégorisée comme une image de bois ; de même pour le kayak, où l'annotation est classée comme « kayak ». Dans le cas où une annotation de bois et de kayak apparaissent simultanément, cette annotation est placée dans un fichier distinct « Bois+Kayak » ; enfin, là où ni le bois ni le kayak n'apparaissent, l'annotation est nécessairement du plastique.

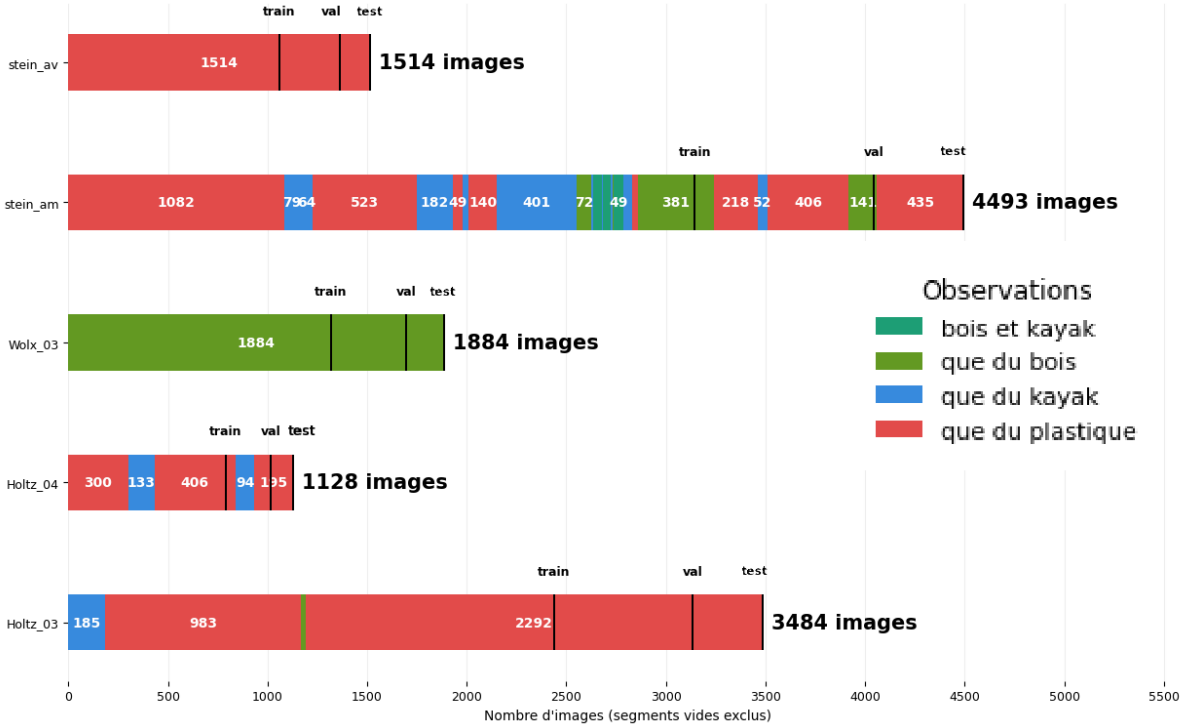


FIGURE 8 – séquence d'observation des annotation sur les vidéo

Cette logique de séparation, appliquée ici pour préserver les classes rares, soulève une seconde question : celle de la stratégie d'entraînement à adopter face à la dégradation de la qualité de la luminosité lors du crépuscule et de la nuit. Les campagnes d'acquisition nocturnes, plus tardives et plus coûteuses à mettre en œuvre, ont produit un volume de données annotées nettement inférieur à celui des campagnes diurnes, ce qui écarte d'emblée un entraînement nocturne indépendant depuis des poids génériques, insuffisamment représentatif pour permettre au réseau de converger correctement. Par ailleurs, la perspective d'un déploiement embarqué sur des ressources limitées invite à interroger la pertinence de maintenir deux modèles spécialisés, l'un diurne et l'autre nocturne, plutôt qu'un modèle unique couvrant les deux régimes : la première solution multiplie l'empreinte mémoire et la complexité de changement entre les modèles sur le dispositif, tandis que la seconde suppose que le réseau parvienne à généraliser des caractéristiques visuelles très différentes sans que l'une des deux conditions ne soit pénalisée par l'autre. Trois configurations d'entraînement sont réaliser : le modèle jour, issu du split temporel décrit précédemment, un modèle nocturne obtenu par affinage des poids par un *fine-tuning* de ce dernier, et un modèle mixte entraîné directement sur l'ensemble des données diurnes et nocturnes réunies.

Entraînement d'un modèle nocturne

Le volume de données annotées de nuit étant sensiblement plus restreint que celui disponible de jour, un entraînement complet depuis des poids génériques n'a pas été retenu pour cette configuration. Le modèle nocturne est donc obtenu par *fine-tuning* du modèle jour déjà entraîné : les poids issus de la configuration diurne servent de point de départ, et l'ensemble du réseau est ré-ajusté sur les données nocturnes annotées, selon la même logique de préservation des représentations déjà apprises que celle retenue pour les stratégies d'initialisation présentées en 3.3.1. Cette approche permet de tirer parti des caractéristiques génériques déjà extraites par le modèle jour tout en spécialisant le réseau sur les conditions de faible luminosité, sans nécessiter un volume de données nocturnes comparable à celui du jeu diurne.

Entraînement d'un modèle mixte

Une seconde configuration, dite mixte, a été entraînée elle directement sur l'ensemble combiné des données jour et nuit disponibles sur les sept échantillons du dispositif, sans étape de *fine-tuning* intermédiaire ni séparation préalable selon la période d'acquisition. L'entraînement est effectuer de la même manière que pour le modèle de jour, en reposant sur l'architecture YOLO26s sur les poids génériques pré-entraînés. Cette configuration mixte vise à évaluer si un modèle unique, exposé aux deux type de luminosité, RGB et infrarouge, dès l'entraînement, peut se substituer à l'association d'un modèle jour et d'un modèle nocturne spécialisé, dans une perspective de simplification du déploiement embarqué.

3.4 Chaîne de traitement

3.4.1 Embarqué

À l'issue de la phase d'entraînement et de validation, l'intégration opérationnelle des modèles sur le site de surveillance environnementale de Steingiessen a imposé quelques contraintes techniques à prendre en compte.

Le déploiement du modèle sur des stations autonomes a en effet exclu le transfert en continu de flux vidéo haute définition par réseau 4G/LTE, en raison de la consommation excessive d'énergie et de la dépendance à la couverture réseau. Le modèle d'inférence doit donc tourner localement et en temps réel directement sur le micro-ordinateur de la station, afin d'extraire des détections les métadonnées qui sont plus légères lors du transfert. Ces métadonnées seront alors uniquement accompagnées de leur image correspondante, afin de valider la détection et de garder un envoi léger.

Le passage d'un modèle d'apprentissage PyTorch de référence (‘.pt‘) à un modèle optimisé pour la cible embarquée nécessite une phase de conversion et de compilation vers des formats d'inférence spécialisés. Le modèle d'origine, bien qu'idéal pour l'entraînement, présente une empreinte mémoire trop lourde et une latence excessive sur les processeurs de faible puissance.Afin d'y remédier, plusieurs formats d'inférence ont été mis en concurrence : le format universel ONNX, le moteur OpenVINO développé par Intel et accompagné de son propre format, et la bibliothèque NCNN conçue par Tencent.

Chacun de ces frameworks répond à des besoins d'optimisation spécifiques. ONNX fait office de format pivot universel facilitant l'interopérabilité entre outils. OpenVINO optimise l'exécution sur les architectures matérielles Intel (x86_64, processeurs graphiques intégrés et accélérateurs de type NPU) en tirant parti d'instructions vectorielles dédiées. NCNN, à l'inverse, est spécifiquement architecturé pour l'inférence sur processeurs embarqués et mobiles de type ARM, en exploitant des noyaux d'exécution optimisés en assembleur (ARM NEON).

Le choix du format le plus adapté à notre application a reposé sur deux critères. Le premier concerne la précision de détection : la précision moyenne (mAP) a été mesurée sur les jeux de données de validation, pour vérifier que l'exportation et la quantification ne dégradaient pas la qualité des détections. La demi-précision FP16 a été retenue, à l'inverse de la quantification INT8, écartée car elle faisait chuter le rappel sur les petits débris. Le second critère porte sur la latence d'inférence et la charge des ressources système, mesurées et comparées pour chaque format. Cette évaluation a d'abord été menée en laboratoire, à l'aide d'un script de monitoring dédié permettant de comparer plusieurs modèles de façon standardisée. Des tests en conditions réelles ont ensuite été réalisés directement sur la cible embarquée : une Raspberry Pi 5, équipée d'un processeur quad-core ARM Cortex-A76 cadencé à 2,4 GHz et d'un capteur Sony IMX708.

3.4.2 Tracking

La détection instantanée par CNN ne suffit pas à quantifier les flux de macro-déchets plastiques : considérer chaque boîte englobante détectée trame par trame conduirait en effet à un surcomptage massif des objets stationnaires ou dérivants. Un algorithme de suivi multi-objets (MOT, Multi-Object Tracking) est donc nécessaire pour relier les détections successives d'un même objet sous un identifiant unique (Track ID). Cette association conditionne directement l'estimation de masse : la reconstruction tridimensionnelle repose sur une analyse statistique de l'objet tout au long de sa trajectoire, afin de repérer l'instant où son grand axe est maximal et son épaisseur minimale, ce qui suppose un suivi continu, sans coupure.

Le milieu fluvial impose cependant des contraintes qui perturbent les algorithmes de suivi classiques. Les variations de luminosité et les reflets du soleil génèrent des détections intermittentes, fragmentant les trajectoires. Les herbiers et branchages des berges, bien que statiques, oscillent sous l'effet du courant et trompent le détecteur YOLO, créant des faux positifs récurrents. Les turbulences et vagues, enfin, peuvent masquer un objet sur plusieurs trames, ce qui impose un mécanisme de prédiction spatiale pour maintenir la trajectoire en l'absence de mesure directe.

Trois approches de suivi temporel ont été comparées dans le cadre de cette étude. Les deux premières sont des algorithmes standards de l'état de l'art : ByteTrack et BoT-SORT. ByteTrack fonde son association sur l'IoU (Intersection over Union) et exploite aussi bien les détections à fort score de confiance que celles à faible score, souvent dues à une occlusion partielle ou à un flou de bougé, ce qui renforce la résilience du suivi. BoT-SORT reprend ce principe et y ajoute une compensation des mouvements de caméra (GMC) ainsi qu'une ré identification des objets par leur apparence (Re-ID) ; cette dernière fonctionnalité se révèle toutefois peu discriminante face aux déchets plastiques de formes similaires et changeantes. La troisième approche est notre traqueur personnalisé, le Geometric_Tracker.

Il a été développé pour dépasser les limites des méthodes génériques, en intégrant des lois physiques et des contraintes propres au contexte fluvial strasbourgeois. Contrairement au filtre de Kalman classique, qui suppose un mouvement ou une accélération constante en deux dimensions, notre traqueur part d'un modèle de mouvement physique connu à l'avance. Une vitesse de dérive descendante par défaut ($V_x = -0{,}02$, $V_y = 0{,}0$, en coordonnées normalisées) est appliquée dès l'initialisation d'une piste, pour refléter le sens d'écoulement de la rivière sur l'image. Cette vitesse est ensuite réactualisée pour chaque piste active, à partir de la médiane glissante des déplacements réels validés sur les 15 dernières images. Ce lissage par la médiane rend le suivi robuste face aux détections aberrantes, qui provoquent des sauts de boîtes englobantes. Deuxièmement, le traqueur intègre un module de filtrage géographique et cinématique des fausses détections statiques, l'Algae Filter. Une zone de forte turbulence et d'accumulation d'herbiers a été cartographiée dans le quart inférieur gauche de l'image ($x \leq 0{,}35$ et $y \geq 0{,}65$). Toute nouvelle piste qui débute dans cette zone est soumise à une règle stricte : si le déplacement horizontal cumulé de son centre reste inférieur à 0,03 après 5 trames, la piste est classée comme bruit statique, de la végétation, et immédiatement écartée du pipeline d'estimation de masse. Troisièmement, la fonction d'association du tracker remplace l'IoU classique, trop sensible aux petites variations de taille sur les petits objets, par un score hybride : 60 % pour la proximité euclidienne des centres, 40 % pour la similarité des surfaces des boîtes. Un objet ne pourras, par conséquent, être associé qu'à une détection proche spatialement et de dimensions cohérentes avec son historique. Enfin, les occlusions sont gérées en maintenant les pistes perdues sous forme de pistes virtuelles, pendant 15 images maximum. Leur position est extrapolée à chaque trame, en prolongeant la trajectoire selon le vecteur de vitesse médiane calculé avant la perte de détection. Leur score de confiance décroît alors de façon exponentielle, jusqu'à suppression définitive ou ré association.

3.5 Estimation des dimensions réelles

Une fois le modèle entraîné, chaque détection de boîte englobante qui est exprimée en pixels doit être convertie en une estimation métrique de la taille réelle de l'objet flottant. Cette conversion reprend le principe de projection géométrique monoculaire proposé par Grimmer et al., 2025, adapté au contexte de caméras de berge fixes. Sa méthode modélise la caméra comme un système sténopé dont les paramètres intrinsèques sont déduits des caractéristiques physiques du capteur, à reconstruire pour chaque détection le faisceau de rayons optiques correspondant aux quatre coins et au centre de la boîte englobante, puis à projeter ce faisceau sur le plan de la surface de l'eau afin d'estimer la position 3D de l'objet et, par différence de distances entre plans opposés, ses dimensions réelles en centimètres. Les valeurs brutes obtenues sont ensuite corrigées par un modèle de régression publié par les mêmes auteurs et calibré sur un jeu d'objets de référence mesurés manuellement. L'ensemble de la chaîne de traitement, depuis la détection jusqu'à l'estimation finale de la masse, est synthétisé en figure 9 ; les deux étapes détaillées ci-après, module d'estimation 3D et application du correctif, en constituent le cœur géométrique.

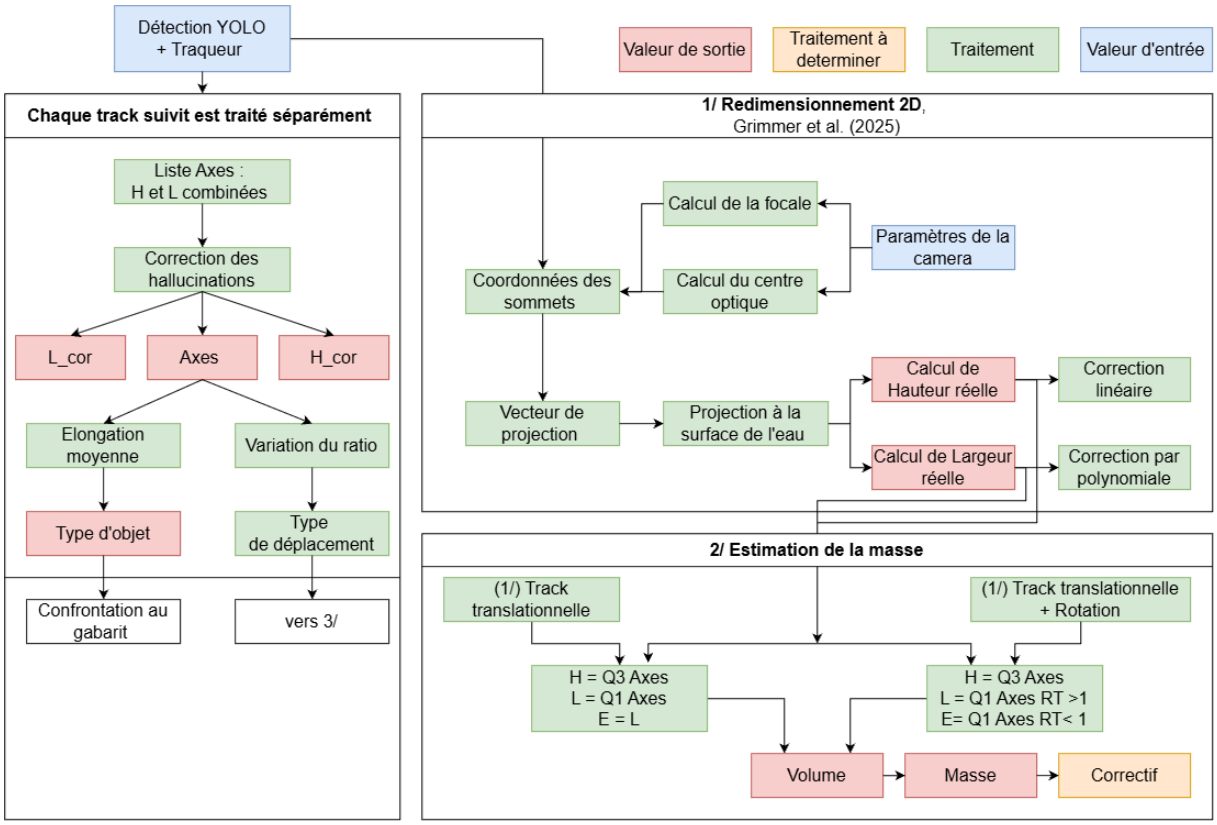


FIGURE 9 – Chaîne de traitement complète de l'estimation de la masse, depuis la détection YOLO en sortie du trackeur et le suivi jusqu'à l'application du facteur correctif final.

3.5.1 Module d'estimation 3D

Enjeux et principe général

La quantification des flux de macro-déchets plastiques dans les hydrosystèmes fluviaux constitue un enjeu opérationnel pour les gestionnaires de bassins versants et les organismes de protection environnementale. Si la détection et le simple comptage d'objets dérivants apportent une première estimation de la charge polluante, ils demeurent insuffisants pour évaluer l'impact environnemental réel ou dimensionner les dispositifs de rétention. La grandeur physique recherchée est la masse cumulée transitant vers l'aval, mais son accès direct par des méthodes manuelles (campagnes de pêche, pesées après collecte) présente un coût logistique prohibitif et une résolution temporelle trop discontinue pour rendre compte de la dynamique des crues, vecteurs majeurs de ces flux. Le recours à un système de vision par ordinateur autonome, capable d'estimer en continu les dimensions métriques, le volume puis la masse des objets à partir d'un flux vidéo monoculaire fixe, constitue à ce titre une alternative pérenne, à condition de s'affranchir de toute mire de calibration complexe ou d'intervention humaine récurrente dans le lit du cours d'eau.

L'estimation dimensionnelle s'appuie sur le formalisme géométrique développé par GRIMMER et al., 2025, fondé sur un modèle de caméra sténopé (*pinhole*) appliqué à un plan d'eau horizontal de référence.

Pour une caméra de focale nominale f, de capteur de dimensions physiques (w_s, h_s), traitant des images de résolution $W_{\text{res}} \times H_{\text{res}}$, les champs de vision horizontal et vertical s'écrivent :

$$\text{FOV}_x = 2\arctan\left(\frac{w_s}{2f}\right), \qquad \text{FOV}_y = 2\arctan\left(\frac{h_s}{2f}\right)$$

Les focales exprimées en pixels ainsi que le centre optique de l'image en découlent :

$$f_x = \frac{W_{\text{res}}}{2\tan(\text{FOV}_x/2)} = f \cdot \frac{W_{\text{res}}}{w_s}, \qquad f_y = \frac{H_{\text{res}}}{2\tan(\text{FOV}_y/2)} = f \cdot \frac{H_{\text{res}}}{h_s}$$

$$c_x = \frac{W_{\text{res}} - 1}{2}, \qquad c_y = \frac{H_{\text{res}} - 1}{2}$$

Pour un pixel d'intérêt (x, y), la direction unitaire du rayon optique associé, exprimée dans le repère caméra, est donnée par :

$$\mathbf{v}_{\text{local}} = \text{normalise}\left(\left[\frac{x - c_x}{f_x},\ \frac{c_y - y}{f_y},\ -1\right]^\top\right)$$

Ces calculs sont réalisés dans le référentiel de résolution native de la caméra (1920×1080 pour les caméras Steingiessen), indépendamment de la résolution d'entrée réduite utilisée pour l'entraînement du réseau de détection (tableau 5) les deux résolutions ne doivent pas être confondues.

Le passage de ce vecteur au repère monde s'effectue par application de la matrice de rotation associée au tangage φ de la caméra, le roulis étant supposé nul et le lacet fixé le long de l'axe d'observation :

$$R_x(\varphi) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\varphi & -\sin\varphi \\ 0 & \sin\varphi & \cos\varphi \end{pmatrix}, \qquad \mathbf{v}_{\text{world}} = R_x(\varphi)\,\mathbf{v}_{\text{local}}$$

Projection et intersection avec le plan d'eau

Une approche naïve consisterait à projeter directement les sommets de la boîte englobante 2D sur le plan de l'eau. Cette méthode introduit toutefois une erreur géométrique systématique, les sommets supérieurs d'un objet flottant correspondant à sa partie émergée et ne touchant pas la surface de l'eau. Pour s'en affranchir, la méthode de Grimmer et al., 2025 construit quatre plans sécants issus du centre optique C et s'appuyant sur les quatre coins de la boîte englobante, dont les directions sont notées $\mathbf{v}_{tl}$, $\mathbf{v}_{tr}$, $\mathbf{v}_{bl}$ et $\mathbf{v}_{br}$:

$$\begin{aligned} \Pi_{\text{top}} &= \text{Plan}(C, \mathbf{v}_{tl}, \mathbf{v}_{tr}) \\ \Pi_{\text{bottom}} &= \text{Plan}(C, \mathbf{v}_{bl}, \mathbf{v}_{br}) \\ \Pi_{\text{left}} &= \text{Plan}(C, \mathbf{v}_{tl}, \mathbf{v}_{bl}) \\ \Pi_{\text{right}} &= \text{Plan}(C, \mathbf{v}_{tr}, \mathbf{v}_{br}) \end{aligned}$$

La position moyenne de l'objet, notée O, est obtenue par intersection du rayon optique central $\mathbf{v}_{\text{center}}$ avec le plan hydrodynamique de référence ($Y = 0$). Les dimensions apparentes projetées de l'objet, largeur L_{proj} et hauteur H_{proj}, correspondent alors à la somme des distances orthogonales de O aux plans qui lui sont opposés :

$$L_{\text{proj}} = \text{dist}(O, \Pi_{\text{left}}) + \text{dist}(O, \Pi_{\text{right}}), \qquad H_{\text{proj}} = \text{dist}(O, \Pi_{\text{top}}) + \text{dist}(O, \Pi_{\text{bottom}})$$

Filtrage des valeurs aberrantes

Les approximations inhérentes à la projection, ainsi que les perturbations visuelles telles que les reflets, le clapotis ou l'instabilité des contours détectés par YOLO, rendent nécessaire un traitement statistique portant sur l'ensemble de la trajectoire de suivi (*track*) de chaque objet, plutôt que sur une détection isolée.

Pour chaque piste de suivi, les projections successives L_{proj} et H_{proj} sont ordonnées, et un filtrage des sauts (*gap filtering*) est appliqué afin qu'une fausse détection ponctuelle ne fausse pas la distribution retenue. L'écart-type de référence σ est calculé sur l'ensemble de la série ; tout écart entre deux mesures consécutives situées aux extrémités de la distribution (sous le premier quartile Q_1 ou au-dessus du troisième quartile Q_3) et dépassant le seuil dynamique

$$\Delta_{\max} = \max(1{,}5\,\sigma,\ 2{,}5 \text{ cm})$$

est identifié comme une aberration géométrique. Les valeurs concernées sont éliminées de manière itérative, sous réserve de conserver un effectif minimal de $N \geq 8$ trames afin de garantir la robustesse statistique du suivi.

Classification par élongation

Les propriétés volumétriques d'un déchet, ainsi que les biais de projection qui lui sont associés, varient fortement selon son gabarit. Chaque objet est par conséquent classifié à l'aide d'un ratio d'élongation robuste E_s, calculé sur les quartiles extrêmes des séries corrigées :

$$E_s = \frac{\overline{L}_{\text{top 25\%}}}{\overline{H}_{\text{bottom 25\%}}}$$

Un seuil $E_s \geq 2{,}2$ (seuil strict retenu à 2,5) conduit à classer l'objet comme une bouteille ou une canette, corps élancé caractérisé par une longueur nettement supérieure à sa largeur apparente ; au sein de cette catégorie, un objet dont $\overline{L}_{\text{top 25\%}} \geq 25{,}0$ cm est distingué comme une grande bouteille, les autres cas relevant de la petite bouteille ou de la canette. À l'inverse, un ratio $E_s < 2{,}2$ caractérise un corps trapu de type bidon, de forme davantage parallélépipédique.

Analyse de la dynamique : translation ou rotation

Le comportement de dérive d'un objet modifie l'exposition de ses différentes faces à la caméra, et conditionne donc la fiabilité des dimensions mesurées. Deux cas sont distingués. Dans le premier cas, dit de translation pure, l'objet conserve une orientation stable par rapport à l'axe optique de la caméra, et la projection transversale mesurée correspond à une dimension constante de l'objet. Dans le second cas, dit de rotation, l'objet est soumis à un mouvement de lacet ou de roulis hydrodynamique qui se traduit par une inversion périodique ou une fluctuation significative du ratio d'aspect de sa boîte englobante, caractérisée par un écart $\text{ratio}_{\max} - \text{ratio}_{\min} > 1{,}25$. Cette distinction dynamique détermine la manière dont les trois axes physiques de l'objet sont ensuite reconstitués à partir des séries de mesures disponibles.

3.5.2 Application de correctif

Modèles de correction géométrique

Les dimensions projetées L_{proj} et H_{proj} sous-estiment ou sur-estiment systématiquement les dimensions réelles de l'objet, en raison des approximations de la méthode de projection et des limites de résolution de la détection. Les régressions empiriques publiées par Grimmer et al., 2025 sont appliquées afin de corriger cet écart pour les corps élancés (bouteilles et canettes), complétées par des gardes-fous linéaires destinés à neutraliser les branches décroissantes des paraboles de régression aux valeurs extrêmes. Pour la longueur, la correction s'écrit :

$$L_{\text{shape}} = 0{,}88\, L_{\text{proj}} + 1{,}29, \qquad L_{\text{finale}} = -0{,}069\, L_{\text{shape}}^2 + 4{,}34\, L_{\text{shape}} - 35{,}94$$

et pour la hauteur :

$$H_{\text{shape}} = -0{,}080\, H_{\text{proj}}^2 + 2{,}38\, H_{\text{proj}} - 5{,}75, \qquad H_{\text{finale}} = 0{,}62\, H_{\text{shape}} + 2{,}98$$

Au-delà de $L_{\text{shape}} > 31{,}45$ cm ou $H_{\text{proj}} > 14{,}875$ cm, la branche parabolique devient décroissante et cesse d'être physiquement cohérente ; une extrapolation linéaire de pente 0,85 est alors substituée, sous un plafond fixé à $L \leq 36$ cm et $H \leq 12$ cm, et un plancher de 3 cm. Pour les bidons, dont la géométrie plus trapue rend la régression parabolique moins pertinente, un modèle linéaire conservateur est retenu :

$$L_{\text{finale}} = \min(1{,}02 \times L_{\text{proj}},\ 36{,}0 \text{ cm}), \qquad H_{\text{finale}} = \min(0{,}90 \times H_{\text{proj}},\ 16{,}0 \text{ cm})$$

avec un plafond dur de 36×16 cm et un plancher de 5 cm.

Extraction des trois axes physiques

À partir des séries corrigées L_{finale} et H_{finale}, les trois dimensions physiques de l'objet, longueur L, largeur l et épaisseur e, sont dérivées selon le cas dynamique identifié précédemment. Pour une bouteille en translation stable (cas A), la longueur retient le troisième quartile de la série corrigée, tandis que la largeur et l'épaisseur, non discriminables faute de rotation de l'objet devant la caméra, sont toutes deux assimilées au premier quartile de la hauteur corrigée : $L = Q_3(L_{\text{finale}})$, $l = e = Q_1(H_{\text{finale}})$. Pour une bouteille en rotation (cas B), la rotation de l'objet permet d'observer ses différentes faces, et la largeur peut alors être distinguée de l'épaisseur : $L = Q_3(L_{\text{finale}})$, $l = Q_3(H_{\text{finale}})$, $e = Q_1(H_{\text{finale}})$. Pour les bidons, une contrainte volumétrique proportionnelle standard, de rapport $1 : 2/3 : 1/3$ entre longueur, largeur et épaisseur, est appliquée et adaptée aux percentiles mesurés lorsque des phases de rotation permettent de séparer effectivement la largeur de l'épaisseur.

3.5.3 Estimation du volume et de la masse

Formulations volumiques

La transformation des dimensions métriques en masse solide s'appuie sur une modélisation volumique par primitive géométrique, pondérée par un coefficient de forme f propre à chaque catégorie d'objet, puis convertie via la masse volumique apparente du matériau.

Pour les bouteilles et canettes, le volume est assimilé à celui d'un cylindre ajusté, dont le diamètre moyen est dérivé des deux axes transversaux :

$$V_{\text{bouteille}} = f \cdot \pi \cdot \left(\frac{l+e}{4}\right)^2 \cdot L$$

Le facteur de forme f est fixé à 0,85 pour les petites bouteilles et canettes, dont le corps s'approche d'un cylindre quasi parfait, et réduit à une plage de 0,65 à 0,70 pour les grandes bouteilles, afin de compenser la présence du col, du goulot et de l'étranglement de prise en main, qui réduisent le volume effectif par rapport à un cylindre plein de mêmes dimensions extérieures.

Pour les bidons, le volume brut du parallélépipède ($L \cdot l \cdot e$) subit un abattement rendant compte de la poignée creuse, de l'espace vide sous le bouchon et des formes biseautées propres à ce type de contenant :

$$V_{\text{bidon}} = f_{\text{bidon}} \cdot (L \cdot l \cdot e), \qquad f_{\text{bidon}} \approx 0{,}75 \text{ à } 0{,}85$$

Conversion en masse

La masse théorique M de l'objet est obtenue par le produit du volume reconstruit V et de la densité apparente ρ_{mat} du matériau :

$$M = V \cdot \rho_{\text{mat}}$$

Pour les corps creux fermés dérivant en surface, tels que les bouteilles en PET ou les bidons en PEHD vides ou partiellement remplis d'air, la densité matière équivalente de l'enveloppe plastique, rapportée au volume total apparent de l'objet, se situe dans une plage standard :

$$\rho_{\text{mat}} \in [0{,}024,\ 0{,}0356] \text{ g/cm}^3$$

3.5.4 Sécurité géométrique et score de confiance

Règles de sécurité géométrique

Afin de garantir l'autonomie et la robustesse opérationnelle du système, les estimations numériques issues du pipeline géométrique sont soumises à un ensemble de règles de repli. Un plafonnement et un plancher durs assurent le rejet automatique des valeurs non physiques issues de régressions divergentes, avec $L_{\text{finale}} \in [3{,}0,\ 36{,}0]$ cm pour les bouteilles et $L_{\text{finale}} \in [5{,}0,\ 36{,}0]$ cm pour les bidons. En cas d'incohérence dimensionnelle majeure détectée sur un bidon, par exemple un ratio $l/L > 0{,}9$ incompatible avec la morphologie attendue, une règle de repli morphologique injecte les proportions physiques standards sur la base de l'axe le plus stable, telles que $l = \frac{2}{3}L$ et $e = \frac{1}{3}L$. Enfin, un filtre de plausibilité de masse soumet à saturation ou à un malus de cohérence physique toute estimation inférieure à 3 g, considérée comme du bruit résiduel, ou supérieure à 500 g, jugée incompatible avec un déchet plastique unitaire flottant.

Score de confiance cumulatif

Au-delà de la masse estimée elle-même, le système associe à chaque objet un indice de confiance $C_{\text{conf}} \in [0, 1]$, formalisé comme un produit de facteurs d'atténuation et de renforcement rendant compte de la fiabilité de chaque étape du pipeline :

$$C_{\text{conf}} = \overline{P}_{\text{YOLO}} \cdot \omega_{\text{track}} \cdot \omega_{\text{calib}} \cdot \omega_{\text{shape}} \cdot \omega_{\text{dyn}} \cdot \omega_{\text{phys}}$$

La confiance de détection $\overline{P}_{\text{YOLO}}$ correspond à la moyenne des scores de confiance de la boîte englobante sur l'ensemble de la trajectoire. Le facteur de trajectoire ω_{track} pénalise les détections éphémères, de trois trames ou moins ($\omega_{\text{track}} \approx 0{,}5$), et valorise à l'inverse les pistes stables et longues, d'au moins huit trames ($\omega_{\text{track}} = 1{,}0$).

La validité de calibration ω_{calib} correspond à la proportion des trames pour lesquelles le triplet formé par les dimensions projetées et la distance à la caméra reste strictement inclus dans le domaine de calibration empirique, défini par [20,5–33,5] cm × [5,5–10,0] cm pour une distance comprise entre 1 et 14 mètres. La netteté morphologique ω_{shape} est fonction de la marge séparant l'élongation mesurée E_s du seuil de décision ambigu, situé entre 2,2 et 2,5 : un ratio proche de cette frontière entraîne un malus, tandis qu'un ratio franc, supérieur à 4,0 ou inférieur à 1,3, conserve un poids unitaire. La pénalité dynamique ω_{dyn} applique une pondération de 0,85 au cas de translation stable, dans lequel l'épaisseur de l'objet n'est pas directement observée, contre un poids unitaire pour le cas de rotation, qui révèle successivement l'ensemble des facettes de l'objet. Enfin, le facteur de plausibilité physique ω_{phys} décroît de manière exponentielle avec l'écart entre la masse estimée et la courbe de régression standard attendue pour la catégorie d'objet considérée.

Ce formalisme permet de restituer, en complément de l'estimation quantitative continue des flux massiques de plastique, un niveau de certitude statistique associé à chaque mesure, exploitable par les gestionnaires pour filtrer les incertitudes de mesure lors de l'établissement des bilans environnementaux finaux.

4 Résultats

Le chapitre précédent a présenté l'ensemble de la chaîne méthodologique développée pour la détection et la quantification des macro-déchets plastiques en milieu fluvial : constitution du jeu de données et entraînement du modèle de détection, suivi multi-objets, reconstruction géométrique tridimensionnelle et estimation de la masse.

Ce chapitre présente les résultats obtenus à chaque étape de cette chaîne, dans le même ordre que leur description méthodologique : sélection et entraînement du modèle de détection, optimisation pour le déploiement embarqué, suivi multi-objets, puis précision des estimations dimensionnelles et de masse.

4.1 Entraînement et sélection des modèles

Cette section présente les performances de détection obtenues selon les principaux paramètres évalués : la taille d'entrée du réseau, l'échelle du modèle YOLO26, la comparaison avec l'architecture YOLOv8, la capacité de généralisation spatiale mesurée par validation croisée par site, et l'effet d'un découpage chronologique des données.

4.1.1 Impact de la résolution

Les modèles ont été comparés selon deux tailles d'entrée du réseau (*imgsz*), 640 et 1280 px, sur l'architecture YOLO26s (tableau 8). Ces deux configurations reposent toutefois sur des jeux d'entraînement distincts (les Configurations 12 et T3 respectivement) et ne constituent donc pas un ablation strict de l'imgsz seul. La résolution 1280 obtient une précision plus faible (P = 0,781 contre 0,883) mais un rappel et un mAP nettement supérieurs (R = 0,626 contre 0,290 ; mAP50 = 0,786 contre 0,315 ; mAP50-95 = 0,314 contre 0,184), pour une latence GPU proche (2,2 ms contre 2,0 ms).

Cet écart se retrouve quelle que soit l'échelle du réseau ou la configuration d'entraînement testée à 640 px : le mAP50 obtenu à 1280 px (0,786) dépasse l'ensemble des résultats obtenus à 640 px, toutes configurations confondues (de 0,173 à 0,315 selon le tableau 9).

Résolution	Instances	P	R	mAP50	mAP50-95	Latence GPU
640	4662	0,883	0,290	0,315	0,184	2,0 ms
1280	2257	0,781	0,626	0,786	0,314	2,2 ms

Table 8 – Impact de la résolution d'image d'entrée sur les performances de test de YOLO26s (GPU NVIDIA RTX 4000 Ada).

4.1.2 Comparaison de l'échelle du réseau (YOLO26n/s/m)

La comparaison des trois échelles du réseau YOLO26 (tableau 9) montre une précision proche entre les trois variantes (P comprise entre 0,875 et 0,883). Le rappel est maximal pour YOLO26s (R = 0,290, contre 0,240 pour YOLO26n et 0,287 pour YOLO26m), tendance suivie par le mAP50 (0,315 pour YOLO26s, contre 0,300 pour YOLO26n et 0,313 pour YOLO26m). Le mAP50-95 croît en revanche de façon monotone avec la taille du réseau (0,154 ; 0,184 ; 0,195 pour n, s et m respectivement). Le coût de calcul suit cette même progression, de 5,2 à 67,9 GFLOPs, avec une latence GPU passant de 1,0 ms (n) à 4,7 ms (m).

Modèle	Paramètres	GFLOPs	P	R	mAP50	mAP50-95	Latence GPU
YOLO26n	2,38 M	5,2	0,875	0,240	0,300	0,154	1,0 ms
YOLO26s	9,47 M	20,5	0,883	0,290	0,315	0,184	2,0 ms
YOLO26m	20,35 M	67,9	0,883	0,287	0,313	0,195	4,7 ms

TABLE 9 – Comparaison des performances de test selon la taille du réseau YOLO26 (GPU NVIDIA RTX 4000 Ada).

Config.	Stratégie	Jeu	P	R	AP_1	AP_2	AP_3	mAP50	mAP50-95
Réf.	Grimmer et al., 2025 (Train 3, YOLOv8n)	—	0,970	0,888	—	—	—	0,939	0,672
1	Initialisation aléatoire	D2	n.d.	[a]	0,914	0,993	0,987	0,965	n.d.
2	Fine-tuning (poids Grimmer et al., 2025)	D2	n.d.	[a]	0,786	n.d.[b]	0,002	0,394[c]	n.d.

TABLE 10 – Comparaison des Configurations 1 et 2 (YOLOv8n, §3.3.1) à la référence publiée par Grimmer et al., 2025. [a] Rappel disponible par classe uniquement (0,90/0,99/0,98 pour la Config. 1 ; 0,84/n.d./0,00 pour la Config. 2). [b] Classe absente du jeu de validation retenu pour cette configuration. [c] mAP50 calculé sur les classes 1 et 3 uniquement ; non directement comparable aux autres valeurs de la colonne.

4.1.3 Méthode statistique

Les Configurations 1 et 2 (§3.3.1), entraînées sur YOLOv8n, sont comparées à la référence publiée par Grimmer et al., 2025 (tableau 10). La Configuration 1 (initialisation aléatoire, jeu D2) atteint un mAP50 de 0,965, légèrement supérieur au mAP50 de 0,939 rapporté par Grimmer et al., 2025, avec un rappel par classe compris entre 0,90 et 0,99. La Configuration 2 (fine-tuning à partir des poids de Grimmer et al., 2025, sur ce même jeu D2) s'effondre au contraire à un mAP50 de 0,394 : la classe 3 chute à un AP de 0,002 et un rappel nul, tandis qu'aucune instance de la classe 2 n'est présente dans le jeu de validation retenu pour cette configuration, ce qui l'exclut du calcul et invite à ne pas comparer directement les deux valeurs de mAP50. La précision globale et le mAP50-95 des Configurations 1 et 2 n'ont pas été extraits à ce stade et ne peuvent donc être confrontés aux 0,970 et 0,672 rapportés par Grimmer et al., 2025.

La Configuration 3 n'a pas pu être incluse dans la comparaison présentée au chapitre 4 : les logs de cet entraînement n'ont pas été conservés, et le temps imparti à l'étude n'a pas permis de le réeffectuer..

4.1.4 Généralisation spatiale (validation croisée par site)

Les effectifs de test propres à chaque site sont rapportés dans le tableau 11. La validation croisée par site montre une forte variabilité des performances selon la composition des jeux d'entraînement et de test (tableau 11, figure 10). Les meilleurs scores sont obtenus lorsque Holtzheim (H3 ou H4) constitue le jeu de test (mAP50 moyen de 0,421 et 0,604 respectivement), et les plus faibles lorsque Steingiessen amont (Sam) constitue le jeu de test (mAP50 moyen de 0,151). Sav et W3 affichent des écarts-types élevés sur la précision (0,270 et 0,222), traduisant une forte dépendance au site d'entraînement associé. Le détail des 20 combinaisons individuelles est présenté en annexe (tableau 11).

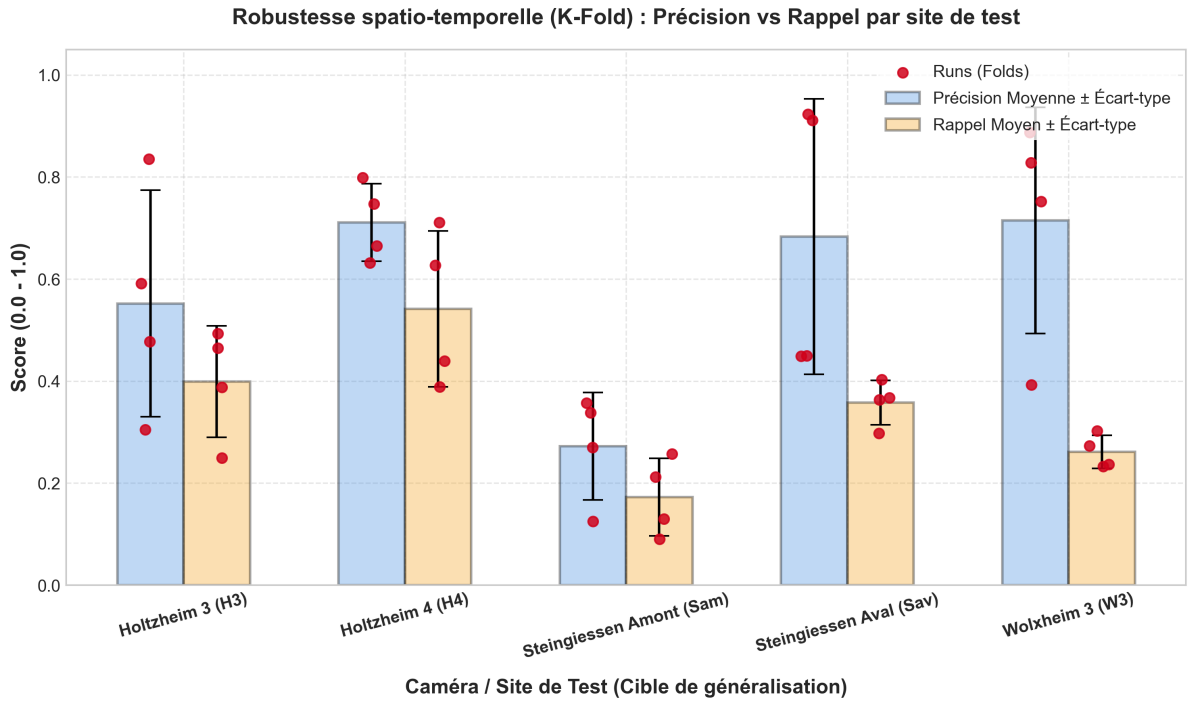


FIGURE 10 – Précision et rappel moyens (± écart-type) obtenus par site de test lors de la validation croisée par site, modèle YOLO26s.

Site de test	Images	Instances	mAP50 (moy. ± ét.)	mAP50-95 (moy. ± ét.)
H3 (Holtzheim, mars)	3509	7303	0,421 ± 0,127	0,212 ± 0,086
H4 (Holtzheim, avril)	1128	2257	0,604 ± 0,169	0,248 ± 0,071
Sam (Steingiessen amont)	4499	14731	0,151 ± 0,090	0,062 ± 0,040
Sav (Steingiessen aval)	1514	2405	0,390 ± 0,033	0,222 ± 0,027
W3 (Wolxheim, mars)	1604	21126	0,277 ± 0,059	0,150 ± 0,029

TABLE 11 – Synthèse de la validation croisée par site : effectifs de test et performance (moyenne ± écart-type sur les 4 combinaisons d'entraînement testées, n=4 par site).

4.1.5 Entraînement par disposition temporelle

Les trois configurations d'entraînement issues du split temporel, jour, nuit et mixte, sont comparées sur leurs jeux de test respectifs (tableau 13).

Le modèle jour obtient les meilleurs scores globaux (mAP50 = 0,877), la classe bois y étant la mieux détectée (mAP50 = 0,965) devant le kayak (0,988 sur un effectif réduit à 140 instances) et le plastique, en retrait (0,678). Testé sur données nocturnes, le modèle nuit obtenu par *fine-tuning* plafonne à un mAP50 de 0,761, la classe kayak, la plus rare des trois avec seulement 16 instances de test, concentrant l'essentiel du déficit (mAP50 = 0,604). Le modèle mixte, entraîné directement sur l'ensemble des sept caméras, obtient sur ces mêmes données nocturnes un mAP50 de 0,836, supérieur au modèle nuit spécialisé sur chacune des trois classes, l'écart étant particulièrement marqué sur le kayak (0,767 contre 0,604).

Modèle	Classe	Images	Instances	P	R	mAP50	mAP50-95
Jour	Toutes	1221	6167	0,959	0,834	0,877	0,662
Jour	Plastique	1221	4031	0,919	0,581	0,678	n.d.
Jour	Bois	613	1996	1,000	0,964	0,965	n.d.
Jour	Kayak	140	140	0,958	0,957	0,988	n.d.
Nuit	Toutes	649	1035	0,768	0,736	0,761	0,475
Nuit	Plastique	111	195	0,852	0,826	0,888	0,533
Nuit	Bois	522	824	0,885	0,694	0,791	0,480
Nuit	Kayak	16	16	0,568	0,688	0,604	0,414
Mixte	Toutes	1865	6900	0,884	0,799	0,836	0,604
Mixte	Plastique	1316	4104	0,861	0,704	0,796	0,442
Mixte	Bois	1151	2747	0,976	0,898	0,944	0,805
Mixte	Kayak	49	49	0,815	0,796	0,767	0,565

Table 12 – Comparaison des modèles de nuit et mixte

Ce résultat, à rebours de l'objectif de spécialisation visé par le *fine-tuning* nocturne, s'explique vraisemblablement par le volume de données sensiblement plus restreint sur lequel repose cette seconde configuration.

Les matrices de confusion normalisées apportent un éclairage complémentaire au tableau 13, en particulier sur la répartition des fausses détections vis-à-vis du fond de l'image. Le modèle jour, malgré ses meilleurs rappels (plastique 0,82, bois 0,98, kayak 0,91), concentre la quasi-totalité de ses faux positifs sur la classe plastique : 98 % des instances de fond faussement détectées lui sont attribuées, contre seulement 1 % pour le bois et 1 % pour le kayak.

Cette asymétrie se retrouve dans les courbes précision-rappel associées (figure 1 13) : la classe plastique conserve la meilleure précision moyenne (AP = 0,888), tandis que le kayak, minoritaire dans le jeu d'entraînement, présente l'AP la plus faible et la plus irrégulière (AP = 0,604), sa courbe présentant plusieurs paliers marqués révélateurs d'un faible nombre d'instances de test. Le mAP@0,5 global s'établit à 0,761 pour ce modèle.

Le modèle mixte conserve un profil de rappel proche (0,82, 0,92 et 0,82 respectivement) et une répartition des faux positifs similaire, quoique légèrement moins marquée (85 % attribués au plastique, 14 % au bois). Les AP par classe suivent toutefois un classement différent de celui observé de jour (figure 3 13) : le bois obtient ici la meilleure précision moyenne (AP = 0,944), suivi du kayak (AP = 0,767) puis du plastique (AP = 0,796), pour un mAP@0,5 global de 0,836, supérieur à celui du modèle jour malgré des rappels comparables ce qui suggère que le gain provient d'une meilleure précision aux seuils de confiance élevés plutôt que d'une meilleure couverture des objets.

Le modèle nuit présente à l'inverse le profil le plus contrasté : un rappel élevé sur la classe plastique (0,88) mais plus modéré sur le bois (0,74) et le kayak (0,75), et surtout une inversion de la répartition des faux positifs, la majorité d'entre eux étant cette fois attribuée au bois (70 %) plutôt qu'au plastique (27 %). Cette inversion suggère que le modèle nuit peine davantage à distinguer les branchages et la végétation en mouvement du bruit de fond nocturne que les modèles jour et mixte ne le font en conditions diurnes, où c'est au contraire le plastique qui concentre l'essentiel des confusions avec l'arrière-plan.

Jour

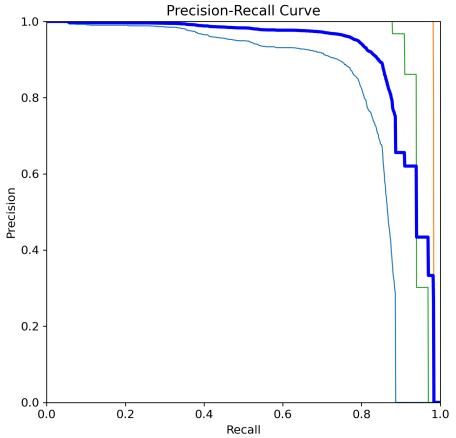
Vrai / Prédit\	plastique	bois	kayak	fond
plastique	0.82	0	0.03	0.98
bois	0	0.98	0	0.01
kayak	0	0	0.91	0.01
fond	0.18	0.02	0.06	

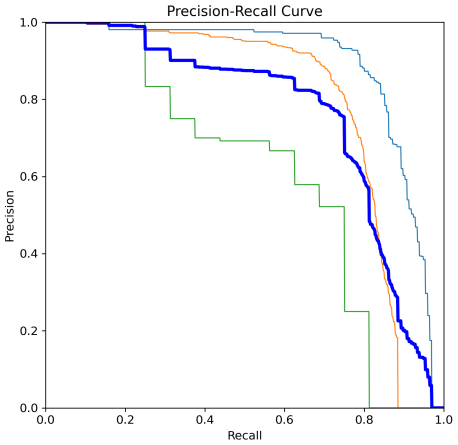
Nuit

Vrai / Prédit\	plastique	bois	kayak	fond
plastique	0.88	0	0	0.27
bois	0	0.74	0	0.70
kayak	0	0.01	0.75	0.03
fond	0.12	0.25	0.25	

Mixte

Vrai / Prédit\	plastique	bois	kayak	fond
plastique	0.82	0	0.02	0.85
bois	0	0.92	0	0.14
kayak	0	0	0.82	0.01
fond	0.18	0.08	0.16	





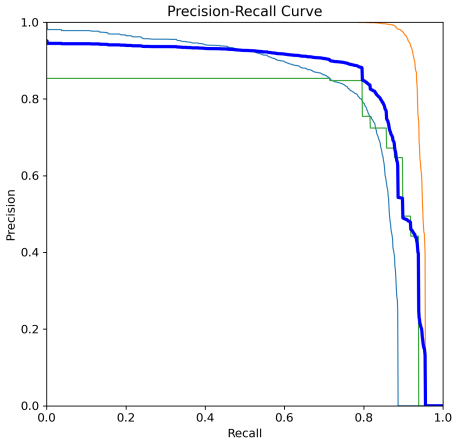


TABLE 13 – Matrices de confusion normalisées par condition d'acquisition (jour, nuit, mixte) / Courbe précision-rappel conditions diurnes, nocturne et mixte

La courbe précision-rappel correspondante (figure 2 13) confirme cette difficulté sur la classe bois (notée 2), dont l'AP (0,791) reste inférieure à celle de la classe plastique (notée 1, AP = 0,888), tandis que la classe kayak (notée 3, AP = 0,604) demeure la plus pénalisée, cohérente avec sa faible représentation dans les annotations nocturnes.

Sur les trois conditions, la classe minoritaire kayak affiche systématiquement l'AP la plus irrégulière, avec des courbes en paliers caractéristiques d'un faible effectif de test, ce qui rejoint le constat du déséquilibre de classes déjà soulevé en amont (60 % plastique, 35 % bois, 3 % kayak) et motive le recours à une pondération accrue de la perte de classification sur cette catégorie.

4.2 Déploiement embarqué

Cette section évalue l'effet de la conversion du modèle retenu vers différents formats d'inférence optimisés pour la cible embarquée, en comparant la stabilité des métriques de détection et les latences associées à chaque format.

La conversion vers les formats d'inférence embarqués (tableau 14) préserve un niveau de précision proche entre les trois formats testés (P entre 0,906 et 0,921 ; mAP50 entre 0,908 et 0,909). OpenVINO se distingue par un poids de modèle réduit (18,59 Mo, comparable à NCNN à 18,25 Mo, contre 36,77 Mo pour ONNX) et par une latence d'inférence nettement inférieure (13,2 ms, contre 63,0 ms pour ONNX et 72,8 ms pour NCNN).

Format	Poids (Mo)	P	R	mAP50	mAP5095	Prétrait.	Inférence	Posttrait.
ONNX	36,77	0,921	0,865	0,908	0,636	0,9 ms	63,0 ms	0,5 ms
OpenVINO	18,59	0,906	0,868	0,909	0,646	0,8 ms	13,2 ms	0,1 ms
NCNN	18,25	0,914	0,872	0,908	0,636	0,9 ms	72,8 ms	0,5 ms

Table 14 – Stabilité métrique et latences comparées après conversion (banc de test : PC de laboratoire).

Format	Latence moyenne (ms)	FPS	Charge CPU (%)	Température CPU (°C)	F
PyTorch (.pt)	795,3	1,25	82,9	50,7	
ONNX (.onnx)	468,0	2,13	98,7	54,5	
NCNN FP16 (.ncnn)	199,5	5,01	89,7	52,4	

Table 15 – Performance sur cible embarquée (Raspberry Pi 5, résolution d'entrée 896×512 px) des trois formats d'inférence pour le modèle YOLO26s retenu en production.

Ces formats, complétés par le modèle PyTorch non converti, ont ensuite été évalués directement sur la cible embarquée (Raspberry Pi 5, résolution d'entrée 896×512 px), pour le modèle YOLO26s retenu en production (tableau 15). Le format OpenVINO n'a pas pu être testé sur cette cible, son moteur d'inférence nécessitant une interface d'exécution dédiée non disponible sur l'instance de Raspberry Pi utilisée. Le format NCNN FP16 obtient la latence la plus faible parmi les formats testés (199,5 ms, soit 5,01 FPS), contre 468,0 ms (2,13 FPS) pour ONNX et 795,3 ms (1,25 FPS) pour le modèle PyTorch non converti. La charge CPU reste élevée pour les trois formats (82,9 à 98,7 %), pour une température CPU comprise entre 50,7 et 54,5 °C. L'usage RAM varie peu selon le format (7,4 à 8,8 %).

4.3 Suivi multi-objets

Trois approches de suivi ont été comparées sur deux séquences de test, Field 1 et Field 2 : notre `Geometric_Tracker`, ByteTrack et BoTSORT (tableau 16). Sur Field 1, le tracker manuel produit un nombre de pistes stables proche du nombre réel d'objets introduits (11 pour 10 objets), tandis que BoT-SORT et ByteTrack en produisent nettement moins (2 à 3). Sur Field 2, les trois trackers dépassent le nombre réel d'objets (13 à 14 pour 10), les deux algorithmes standards davantage que le tracker manuel.

Tracker	Séquence	Détections	Pistes (≥8 frames)	Objets réels
Manuel	Field 1	484	11	10
Manuel	Field 2	814	13	10
BoT-SORT	Field 1	107	3	10
BoT-SORT	Field 2	792	14	10
ByteTrack	Field 1	111	2	10
ByteTrack	Field 2	788	14	10

TABLE 16 – Comparaison des trackers : nombre de pistes stables (au moins 8 frames) au regard du nombre réel d'objets introduits par séquence.

4.4 Estimation des dimensions et de la masse

4.4.1 Limites de la correction polynomiale de Grimmer

Le pipeline en deux étapes de GRIMMER et al., 2025 (correction de forme puis correction de dimension), appliqué avec ses coefficients originaux mais sans mécanisme de sécurité, produit des valeurs aberrantes en dehors de son domaine de calibration initial (bouteilles de 20,5 à 33,5 cm). Sur les 2048 détections appariées à une vérité terrain (tableau 17), les prédictions atteignent jusqu'à -6271,2 cm en largeur et -2063,0 cm en hauteur.

	Largeur		Hauteur	
	MAE (cm)	RMSE (cm)	MAE (cm)	RMSE (cm)
coefficients originaux	143,64	749,95	47,69	181,20
garde-fous + classification	3,93	5,77	2,45	2,55

TABLE 17 – Comparaison globale sur les 2048 détections appariées : pipeline de Grimmer appliqué sans garde-fou vs notre adaptation.

Le détail par catégorie d'objet (tableau 18) montre que cette divergence n'est pas uniforme : les catégories les plus proches du domaine de calibration initial (Grande Bouteille, Bidon) restent dans une gamme d'erreur de quelques dizaines de centimètres, tandis que les objets aux dimensions plus extrêmes (Petite Bouteille, Rocheval) atteignent des MAE de plusieurs centaines à plusieurs milliers de centimètres.

		MAE Largeur (cm)		MAE Hauteur (cm)	
Objet	n	Grimmer brut	Notre adapt.	Grimmer brut	Notre adapt.
Grande Bouteille	1829	12,65	3,44	11,81	2,37
Petite Bouteille	79	2626,35	14,18	668,51	5,30
Bidon 1	40	18,70	6,99	20,94	2,00
Bidon 2	24	11,73	7,00	14,14	2,00
Coca	50	43,45	2,05	15,34	1,00
Rocheval	26	2321,26	2,94	819,48	3,00

TABLE 18 – Erreur absolue moyenne par catégorie d'objet : pipeline de Grimmer sans garde-fou vs notre adaptation (garde-fous, extrapolation linéaire, branche bidon dédiée).

4.4.2 Validation sur données de suivi réelles

Le pipeline complet (projection, classification par élongation E_s, correction, estimation de masse) a été appliqué aux pistes de suivi issues du tracker manuel sur les deux séquences de test réelles, avec vérité terrain disponible (tableaux 11 et 12). La classification par élongation identifie correctement les trois bidons connus de Field 1 (pistes 1, 3 et 7), confirmée par correspondance directe des identifiants de piste avec la vérité terrain.

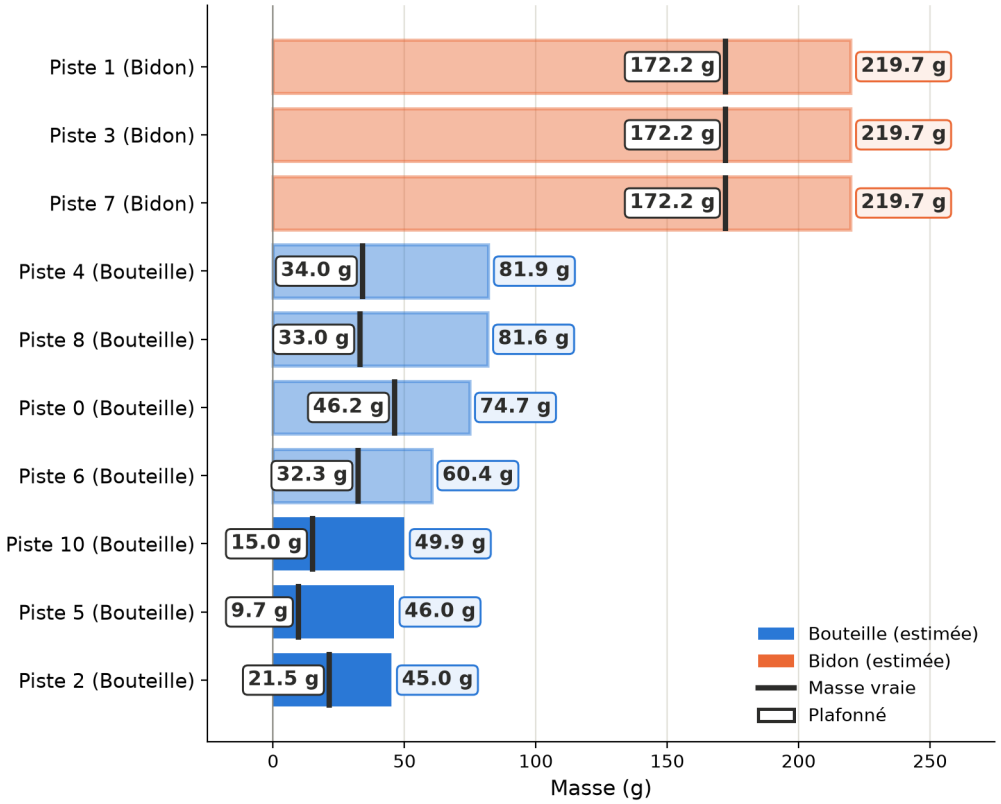


FIGURE 11 – Résultats de masse par piste, Field 1.

4.4.3 Estimation du volume et de la masse

Sur l'ensemble des pistes retenues ($N \geq 8$ trames), la masse moyenne estimée s'élève à 107,3 g par objet sur Field 1 (11 pistes) et à 109,6 g par objet sur Field 2 (13 pistes), toutes catégories confondues. Cette moyenne est tirée vers le haut par les pistes plafonnées : en se limitant aux pistes non plafonnées, la masse moyenne descend à 47,0 g (bouteilles, Field 1) et à 84,4 g (toutes catégories, Field 2), cette dernière valeur restant élevée en raison d'une piste bidon proche du plafond sans l'atteindre (piste 16, 212,4 g). Ces valeurs non plafonnées restent plus proches des masses de référence mesurées manuellement sur les objets utilisés lors de l'acquisition (9,7 g à 46,2 g pour les bouteilles, 167 g à 181 g pour les bidons) que les moyennes globales.

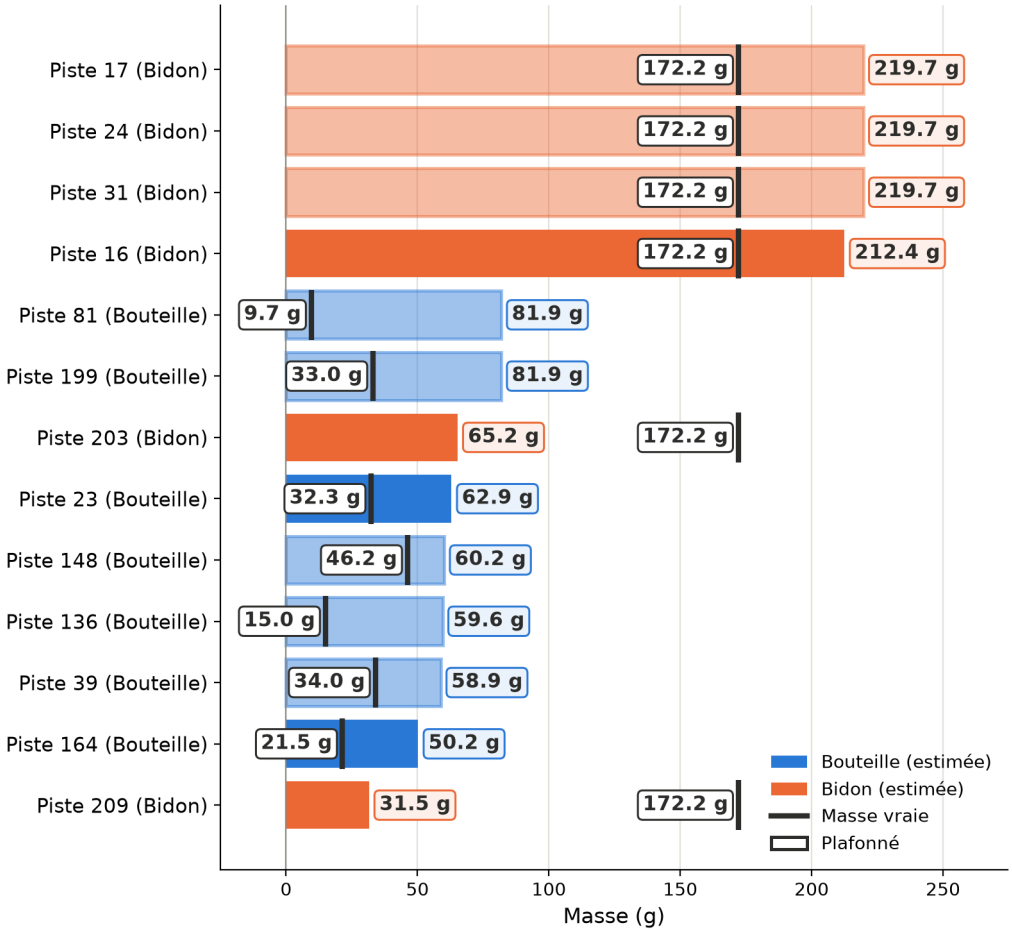


FIGURE 12 – Résultats de masse par piste, Field 2.

5 Discussion

Le précédent chapitre a permis d'exposer les résultats obtenus à chaque étape du développement, les performances du modèle de détection, le comportement du suivi multi-objets, ainsi que la précision des estimations dimensionnelles et de masse. Ce chapitre propose de discuter des limites de ces résultats, ainsi que de les situer par rapport aux travaux existants sur le sujet.

Cette discussion s'organise en trois temps. Les résultats sont d'abord replacés dans leur contexte scientifique, en les comparant aux performances rapportées par les études de référence, notamment celle de Grimmer et al., 2025. Les limites méthodologiques du dispositif sont ensuite examinées, qu'elles relèvent des contraintes du terrain, les choix faits lors de l'entraînement du modèle, ou des approximations inhérentes à la reconstruction géométrique et à l'estimation de masse. Enfin, des perspectives d'améliorations sont proposées.

5.1 Mise en perspective des résultats

5.1.1 Choix de l'architecture et performances

Le choix de YOLO26 a été établi sur l'hypothèse faite par la littérature, que les gains de performance annoncés par Ultralytics, combinés à son orientation vers l'embarqué, en faisaient un candidat idéale pour ce contexte. Toutefois, Grimmer et al., 2025 avait pu observer sur d'anciennes versions de YOLO qu'un modèle plus ancien était en mesure de concurrencer un modèle plus récent, dans son cas, YOLOv8 offrant un meilleur compromis que YOLOv11. Le premier comparatif que nous avons mené entre résolutions et échelles de YOLO26 confirme que les gains de performance ne sont pas uniformes : l'augmentation de la résolution d'entrée (640 vers 1280 px) apporte un gain substantiel sur toutes les configurations testées, tandis que l'augmentation de la taille du réseau (n vers m) n'apporte qu'un gain marginal en performance sur le mAP50 et un léger surcoût en mAP50-95 tout en impactant largement le poid du modèle. Ce constat rejoint celui de Grimmer et al., 2025 : le compromis architecture/performance dépend fortement du contexte, et l'architecture la plus récente n'est pas systématiquement la plus performante pour une tâche donnée.

Nos performances restent nettement inférieures à celles rapportées par Grimmer et al., 2025, qui obtiennent des valeurs de mAP50 comprises entre 0,933 et 0,965 pour leurs variantes YOLOv5 à v11, et des mAP50-95 entre 0,655 et 0,747, contre 0,315 à 0,786 (mAP50) et 0,154 à 0,314 (mAP50-95) pour nos propres configurations. Cet écart n'indique pas nécessairement une architecture ou un entraînement moins performant, il reflète avant tout une différence de protocole d'évaluation. Grimmer et al., 2025 soulignent que la fuite de données temporelle (data leakage) gonfle artificiellement les métriques de plus de 40 points de mAP50 dans leur propre étude lorsque le split n'est pas contrôlé pour l'indépendance temporelle. Notre validation croisée contenant plus de sites est volontairement plus complexe, avec des sites d'entraînement et de test totalement disjoints, elle est donc structurellement plus sévère lors d'une évaluation sur des splits aléatoires classique, ce qui explique une partie significative de l'écart observé.

5.1.2 Méthode statistique

La comparaison à la référence de Grimmer et al., 2025 (tableau 10) doit être lue avec prudence. La Configuration 1 atteint un mAP50 très proche, voire légèrement supérieur (0,965 contre 0,939), mais ce résultat est obtenu sur D2, un split aléatoire 80/20 mêlant l'ensemble des sites sans séparation contrôlée exactement le type de configuration que ce travail identifie par ailleurs (§5.1.2) comme structurellement optimiste par rapport à un split organisé par site. Le léger avantage apparent de la Configuration 1 sur la référence de Grimmer et al., 2025 ne peut donc pas être interprété comme une preuve que l'enrichissement du jeu de données améliore la capacité de généralisation du modèle ; il pourrait tout aussi bien refléter une fuite de données comparable, ou plus importante, que celle affectant potentiellement le protocole d'origine. Seule une évaluation de la Configuration 1 sur un jeu test totalement disjoint des sites d'entraînement permettrait de trancher.

La Configuration 2 pose un problème distinct. Un fine-tuning à partir d'un modèle déjà spécialisé sur le contexte fluvial de Strasbourg, entraîné sur ce même jeu D2 favorable en théorie à la fuite de données, aurait dû au minimum égaler la Configuration 1 ; il s'effondre au contraire à un mAP50 de 0,394. Cet écart oriente vers un défaut de protocole de transfert plutôt que vers un simple manque de données : une incompatibilité entre l'indexation des classes du modèle source (`last.pt` de Grimmer et al., 2025) et celle du jeu D2 est l'hypothèse la plus probable, à vérifier en confrontant directement les fichiers de classes des deux modèles. L'absence totale d'instances de la classe 2 dans le jeu de validation de cette configuration interdit par ailleurs toute conclusion sur cette classe spécifiquement.

Enfin, les identifiants « Configuration 12 » et « T3 » utilisés au tableau 8 pour l'ablation de résolution restent à ce jour non définis dans le texte et gagneraient à être reliés explicitement à la nomenclature des Configurations 1/2 ci-dessus et des combinaisons de validation croisée (tableau 7), afin d'éviter toute ambiguïté sur les runs mobilisés.

5.1.3 Généralisation spatiale et fuite de données

Au-delà du choix de l'architecture, l'écart observé avec Grimmer et al., 2025 invite à examiner le rôle du protocole d'évaluation lui-même dans la performance mesurée. La comparaison entre les métriques de validation obtenues en cours d'entraînement (proches de 0,99 de mAP50) et les métriques de test sur site réellement indépendant (0,151 à 0,604 selon le site) illustre concrètement ce phénomène sur nos propres données. Cet écart massif entre performance apparente et performance réelle confirme, comme le suggèrent Grimmer et al., 2025 et la littérature plus large sur la fuite de données (Babu et al., 2024 ; John et al., 2025), qu'une évaluation sur split aléatoire ou sur données de validation interne surestime substantiellement la capacité de généralisation d'un modèle entraîné sur des séquences vidéo continues.

La forte variabilité observée entre sites de test (mAP50 moyen de 0,151 pour Steingiessen amont contre 0,604 pour Holtzheim avril) et les écarts-types élevés sur certains sites (Sav, W3) suggèrent que la performance du modèle dépend plus de la similarité entre le site d'entraînement et le site de test que de la quantité de données d'entraînement disponible, suggérant dans se cas la présence de data leakage. La sur-performance systématique observée lorsque Holtzheim sert de site de test, et se quel que soit le site d'entraînement associé, est en partie attribuable au fait que les deux sous-ensembles de Holtzheim (mars) et Holtzheim (avril) proviennent de la même caméra et donc du même arrière-plan : leur indépendance n'est donc pas totale au sens strict de la validation croisée par site, ce qui a pu favoriser artificiellement les scores obtenus dans ces configurations.

5.1.4 Effet du split temporel

Ce contraste entre protocoles d'évaluation se retrouve également dans la comparaison entre split temporel et split par site. En effet le split temporel produit des performances légèrement supérieures à la validation croisée par site (mAP50 de 0,836 à 0,877 contre 0,151 à 0,604). Cet écart s'explique par la nature du découpage : le split temporel sépare en effet des images issues de la même caméra à des instants différents, tandis que le split par site sépare des caméras aux conditions d'acquisition distinctes (paysage, angle de vue, luminosité). Le split temporel évalue donc une forme de généralisation restreinte aux changement de conditions sur un site déjà compris par le modèle à la différence du split par site, qui teste la généralisation à un site jamais observé. Les deux protocoles répondent donc à des questions différentes : le premier anticipe la dérive de performance dans le temps sur un site opérationnel tandis que le second anticipe le déploiement sur un nouveau site.

La construction de ce split temporel repose elle-même sur une règle de priorisation par classe (bois > kayak > bois+kayak > plastique, §3.3.3), destinée à préserver la représentation des classes rares dans le jeu de test. Cette règle introduit toutefois un biais potentiel symétrique à celui identifié en §5.1.3 pour la validation croisée par site : en regroupant les annotations par classe avant d'appliquer la frontière temporelle, elle peut recomposer, au sein d'une même classe, des sous-ensembles d'entraînement et de test partageant des séquences vidéo ou des conditions d'acquisition proches, sans que cela soit visible dans le découpage global présenté en figure 8. Cette hypothèse n'a pas été vérifiée à ce stade une vérification directement analogue à celle menée en §5.1.6 pour la correction géométrique de Grimmer et al., 2025, à savoir confronter les identifiants de séquence vidéo effectivement présents de part et d'autre de la frontière temporelle pour chaque classe, permettrait de la confirmer ou de l'écarter.

5.1.5 Absence d'effet de la pondération de la loss

Indépendamment du protocole d'évaluation retenu, plusieurs choix d'entraînement ont également été testés pour tenter d'améliorer ces performances. Parmi les configurations testées, différents paramètres de gestion de la donnée ont été mis en application, la pondération de la loss (*focal loss* ainsi que l'ajustement du *weight decay*). Les entraînements effectués avec ces conditions n'ont produit aucune variation mesurable par rapport à leurs configurations de référence. Ce résultat, à première vue surprenant, pourrait s'expliquer par une gestion automatique de ces mécanismes de régularisation, intégrée directement à YOLO26 contrairement aux versions plus anciennes, ce qui nécessitait leur paramétrage manuel explicite sans effet observable. Cette explication reste toutefois une hypothèse : elle n'a pas été vérifiée directement dans le code source du framework Ultralytics, et une cause technique alternative (paramètre non pris en compte par l'interface en ligne de commande, ou écrasement par des valeurs par défaut) ne peut être exclue à ce stade.

5.1.6 Embarqué

Les mesures obtenues directement sur Raspberry Pi 5 (tableau 15) ne permettent de comparer que trois des formats évalués en laboratoire sur PC (tableau 14) : ONNX, NCNN et PyTorch. OpenVINO pourtant le format le plus rapide en laboratoire, avec 13,2 ms contre 63,0 ms pour ONNX et 72,8 ms pour NCNN - n'a pas pu être testé sur la cible réelle, son moteur d'inférence nécessitant une interface d'exécution dédiée qui n'était pas disponible sur l'instance de Raspberry Pi utilisée.

Cette absence limite la portée de la comparaison embarquée : parmi les formats effectivement testés sur site, NCNN FP16 est le plus rapide (199,5 ms contre 468,0 ms pour ONNX et 795,3 ms pour PyTorch non converti), mais rien ne permet d'affirmer qu'il resterait le meilleur choix si OpenVINO avait pu être inclus dans cette comparaison. Le choix opérationnel du format NCNN FP16 pour le déploiement (§3.4.1) repose donc, à ce stade, sur une comparaison partielle plutôt que sur une confrontation complète des trois formats sur la cible finale.

Des mesures complémentaires, réalisées sur la même cible pour les échelles YOLO26s et YOLO26m, confirment quantitativement le choix de YOLO26s pour le déploiement, déjà motivé en §3.4.1 par des contraintes de latence. En NCNN FP16, la latence passe d'environ 195 ms pour YOLO26s à environ 590 ms pour YOLO26m, soit un facteur proche de 3 ; en PyTorch non converti, YOLO26m atteint 1,7 seconde par image, incompatible avec tout traitement en temps réel sur cette cible. En ONNX, l'écart est nettement plus resserré (environ 405 ms contre 575 ms, facteur $\approx 1{,}4$), un comportement qui contraste avec celui des deux autres formats et qui reste, à ce stade, sans explication.

Ces résultats appellent néanmoins deux réserves. Premièrement, les valeurs de charge CPU et de température rapportées pour NCNN FP16 dans la synthèse officielle (89,7 % ; 52,4 °C) ne coïncident pas avec la moyenne des mesures brutes obtenues pour ce qui semble être la même configuration (45 à 52 % ; 45 à 46 °C) ; cet écart reste à documenter avant toute exploitation de ces chiffres. Deuxièmement, bien que les scripts de journalisation brute aient conservé par défaut l'étiquette générique YOLO11 issue de la version précédente du framework de test, il s'agit bien des poids et de l'architecture YOLO26s/m qui ont été convertis et évalués en inférence sur la Raspberry Pi 5. Ces mesures portent par ailleurs uniquement sur la latence, la charge CPU et la température : la précision de détection sur la cible embarquée elle-même n'a pas été revérifiée indépendamment des mesures obtenues sur PC de laboratoire (tableau 14, §4.2), de sorte que la stabilité du mAP après conversion reste, à ce stade, validée en laboratoire mais non sur site.

5.1.7 Suivi multi-objets

Au-delà de la détection image par image, le comportement du modèle lors du suivi temporel constitue un autre axe d'analyse. Le traqueur géométrique manuel produit, sur Field 1, un nombre de pistes stables proche du nombre réel d'objets (11 pour 10), tandis que ByteTrack et BoT-SORT sous-détectent fortement leur présence (2 à 3 pistes). Cet écart est cohérent avec la conception même de ces deux algorithmes standards, pensés pour des scènes où les objets suivis présentent une apparence stable et distinguable du fond, une hypothèse qui est mise à mal par les reflets et la turbidité de l'eau et qui dégradent la confiance de détection sous le seuil nécessaire à l'initialisation d'un track. Sur Field 2, les trois traceurs dépassent au contraire le nombre réel d'objets, ce qui suggère une fragmentation plus marquée des trajectoires sur cette séquence, sans qu'une cause commune aux trois traqueurs ait été formellement établie.

5.1.8 Robustesse de la correction géométrique de Grimmer et al., 2025

Une fois les objets suivis, c'est la fiabilité de la conversion de leurs dimensions en grandeurs physiques qui doit être interrogée. La confrontation du pipeline de correction de Grimmer et al., 2025 à nos propres données de détection révèle une instabilité mathématique majeure hors de son domaine de calibration initial (bouteilles de 20,5 à 33,5 cm). Appliquée telle quelle, la branche polynomiale de la correction diverge jusqu'à des valeurs non physiques : jusqu'à -6271 cm en largeur et-2063 cm en hauteur. Ce comportement n'est pas une erreur de mise en œuvre.

Il s'agit d'une conséquence attendue d'un ajustement polynomial calibré sur une plage de tailles restreinte : en dehors de cette plage, le terme quadratique domine et produit une divergence non bornée. C'est un phénomène classique de sur-apprentissage d'un modèle paramétrique en dehors de son domaine d'entraînement. Grimmer et al., 2025 reconnaissent eux-mêmes n'avoir validé leur méthode que sur un sous-ensemble restreint d'objets et de distances (1 à 14 m). Ils appellent explicitement à des travaux complémentaires pour évaluer sa robustesse hors de ce cadre, ce que notre analyse apporte concrètement.

Notre adaptation, extrapolation linéaire au-delà du sommet de la parabole, planchers et plafonds physiques durs, et surtout une branche de correction dédiée aux bidons, activée par classification sur le ratio d'élongation E_s, réduit l'erreur moyenne de plusieurs centaines de centimètres à quelques centimètres sur l'ensemble des catégories testées. La nécessité de cette branche dédiée est elle-même illustrée par nos données de suivi réelles (Field 1) : appliquée uniformément sans distinction bouteille/bidon, la correction de Grimmer et al., 2025 ramène systématiquement la hauteur des bidons au plancher de sécurité (3,0 cm), loin des 14 à 18,5 cm mesurés, alors que les catégories bouteille restent correctement estimées. Cela confirme que l'erreur n'est pas due à la géométrie de projection elle-même, commune aux deux approches, mais spécifiquement à l'absence, chez Grimmer et al., 2025, d'une étape de classification morphologique en amont de la correction, une contribution propre à ce travail.

5.2 Limites méthodologiques

5.2.1 Représentativité des données d'acquisition

Si la mise en perspective des résultats précédents éclaire leur portée scientifique, plusieurs limites méthodologiques doivent également être prises en compte pour en apprécier la validité. Les données étrangères recensées dans la littérature (18 bases de données) n'ont pu être mobilisées. La raison principale tient à une distinction insuffisante, dans certaines d'entre elles, entre images de déchets statiques (souvent acquises en laboratoire ou sur berge) et images de débris réellement en dérive sur l'eau. Cette distinction facilite certes la constitution de jeux d'entraînement et de validation « propres ». Mais elle limite la compréhension du modèle une fois déployé sur le terrain : n'ayant vu que des objets dans des conditions idéales, le modèle n'apprend pas à composer avec les conditions climatiques changeantes, ni avec la morphologie de déchets dégradés par un long transit fluvial. Les données acquises par drone posent un problème similaire. Elles priorisent l'apprentissage de paysages variés au détriment des débris eux-mêmes, qui, sur de courtes séquences d'observation, se concentrent surtout sur des objets statiques piégés dans des embâcles. Ces données auraient pu enrichir la compréhension du fond de l'image par le modèle (rivière, berges, décor). Mais leur intégration risquait aussi de diluer la précision sur nos propres données, acquises dans des conditions contrôlées mais réalistes.

Il convient de noter que notre propre protocole d'acquisition partage une partie de cette limite : les débris anthropiques ont été introduits manuellement lors d'une campagne contrôlée, ce qui peut produire un jeu de données visuellement plus « propre » que la réalité des déchets réellement transportés par la rivière (état de dégradation, fragmentation, salissure). Cette limite, déjà relevée par Grimmer et al., 2025 sur leur propre protocole similaire, s'applique donc également à notre étude.

5.2.2 Biais et erreurs sur l'annotation

Au-delà de la représentativité des données elles-mêmes, leur annotation constitue une seconde source potentielle de biais. Une relecture des jeux d'annotation a révélé que certaines séries d'images ont été annotées avec la classe 1 à la place de la classe 0, vraisemblablement en raison d'une confusion d'interface lors de l'utilisation de LabelImg. Bien que cette erreur reste limitée à une partie du corpus, elle constitue une source de bruit dans les jeux d'entraînement et de test dont l'ampleur exacte sur les métriques finales n'a pas pu être quantifiée dans le temps imparti à cette étude.

Un autre biais d'annotation, propre cette fois au site de Wolxheim, concerne la présence quasi permanente de bois flottant statique. La caméra y est positionnée à proximité d'un pont et d'un ouvrage de retenue (seuil), une configuration hydro-morphologique où la courbure du cours d'eau favorise la formation d'embâcles où s'accumule ce bois de façon durable. Ce bois a dû être systématiquement labellisé : laisser ces objets non annotés aurait risqué d'introduire un biais inverse, le modèle pouvant associer l'absence d'annotation sur une texture visuellement proche à l'absence d'objet d'intérêt y compris lorsque du bois se détache de l'embâcle et transite réellement dans le courant, notamment lors de crues ou d'évènements météorologiques marqués, un cas où sa détection resterait pourtant pertinente pour le suivi. Cette nécessité de labelliser une classe non ciblée par l'étude, propre à la configuration hydro-morphologique du site, illustre plus largement la difficulté à constituer un corpus représentatif : aucune donnée externe à notre propre acquisition n'a par ailleurs été mobilisée pour former un jeu de test indépendant, de sorte que la capacité de généralisation du modèle à des environnements fluviaux substantiellement différents (turbidité, végétation, luminosité, infrastructures) de ceux de l'Eurométropole de Strasbourg reste à démontrer au-delà du test qualitatif préliminaire mené sur la Bruche.

Deux pistes pourraient limiter le biais lié au bois de Wolxheim à l'avenir : d'une part, enrichir l'annotation par des séquences acquises lors de différents évènements météorologiques et hydrologiques, en particulier lors des crues où la dynamique des embâcles est la plus active ; d'autre part, restreindre la zone d'analyse par un masque de berge, une piste explorée ponctuellement avec Mask2Former mais n'ayant pas, en l'état, produit de résultats concluants.

Absence de données étrangères pour la validation

Ces biais d'annotation, propres à notre propre corpus, posent en creux la question plus large de sa représentativité vis-à-vis d'autres contextes fluviaux. Aucune donnée externe à notre propre acquisition n'a été mobilisée pour constituer un jeu de test indépendant. La capacité de généralisation du modèle à des environnements fluviaux substantiellement différents (turbidité, végétation, luminosité, infrastructures) de ceux de l'Eurométropole de Strasbourg reste donc à démontrer au-delà du test qualitatif préliminaire mené sur la Bruche.

5.2.3 Limites de l'estimation dimensionnelle et de masse

Ces limites concernant les données ne sont pas les seules à peser sur la validité des résultats : la chaîne d'estimation dimensionnelle et de masse comporte elle-même ses propres zones d'incertitude. Notre adaptation de la correction géométrique reste limitée aux catégories d'objets identifiées par le ratio d'élongation E_s (bouteilles, canettes, bidons), et ne couvre pas l'ensemble des morphologies de déchets plastiques rencontrées en rivière, en particulier les sachets et films plastiques, dont la géométrie non rigide et variable ne se prête pas à une classification par élongation simple.

L'absence, dans la littérature comme dans les données publiquement disponibles, d'une table de correspondance fiable entre référence commerciale d'un contenant et sa masse déclarée par le fabricant, limite également la possibilité de valider ou d'affiner les coefficients de densité apparente ρ_{mat} retenus, qui reposent sur une plage standard plutôt que sur une mesure exhaustive par produit.

5.2.4 Limites de la validation sur données réelles

Ces limites théoriques se traduisent concrètement lors de l'application de la méthode aux pistes de suivi réelles. Les résultats obtenus sur Field 1 et Field 2 doivent être interprétés avec prudence. Bien que la classification par élongation se révèle fiable, elle identifie correctement les trois bidons connus de Field 1 par recoupement avec la vérité terrain, Ces résultats mettent en évidence les limites structurelles d'une rétro-projection monoculaire sténopé en milieu fluvial agité, où le clapotis et les variations de flottaison induisent des instabilités de rayons optiques. La saturation observée (73 % et 62 % des pistes atteignant le plafond de garde-fou) démontre que le filtrage statistique a posteriori ne suffit pas : pour transformer ce pipeline en instrument métrologique fiable, un filtrage cinématique et géométrique frame-by-frame en amont de l'agrégation de la piste est indispensable.

Cette saturation ne provient pas des formules de correction elles-mêmes, dont la fidélité aux coefficients de Grimmer et al., 2025 a été vérifiée indépendamment à plusieurs reprises. Elle provient très probablement de la projection géométrique brute (avant correction), dont une part substantielle des détections individuelles, jusqu'à 70 % sur certains lots testés en amont de cette validation, affiche des dimensions projetées largement supérieures à toute taille physique plausible. Deux pistes ont été explorées sans aboutir à une explication complète : une incohérence de résolution entre le référentiel pixel du traqueur et celui du modèle de caméra (partiellement corrigée, sans résoudre le problème en totalité), et une éventuelle instabilité géométrique de l'intersection rayon-plan pour les détections proches de l'horizon de l'image, cette dernière hypothèse n'a pas été confirmée, la distance à la caméra ne se distinguant pas significativement entre détections aberrantes et détections correctes.

Ce constat appelle plusieurs pistes de correction, par ordre de priorité. D'abord, le filtrage des sauts ($\Delta_{\max}$) déjà prévu par la méthode devrait être appliqué au niveau de la trame individuelle avant l'agrégation par piste, plutôt que sur la seule série corrigée finale : cela écarterait les détections isolées aberrantes avant qu'elles ne contaminent les quartiles (Q_1, Q_3) utilisés pour la reconstruction dimensionnelle, plutôt que de laisser le plafond physique masquer leur effet a posteriori. Ensuite, une vérification frame par frame du calcul de projection (vecteurs de rayons, distance à la caméra, orientation) sur un sous-échantillon contrôlé, par exemple les trames correspondant aux trois bidons de Field 1, dont la position réelle est connue, permettrait d'isoler précisément l'origine de la surestimation plutôt que de la traiter statistiquement. Enfin, le score de confiance cumulatif C_{conf}, déjà formalisé dans la méthode mais non encore appliqué à ce stade de la validation, offrirait un moyen direct de signaler et d'écarter les pistes les moins fiables plutôt que de leur attribuer une masse plafonnée arbitrairement identique à d'autres pistes.

5.3 Perspectives d'amélioration

Face à ces limites, plusieurs pistes peuvent être envisagées pour consolider ou prolonger ce travail.

Vers une estimation de masse générique. Une piste de simplification explorée sans être mise en œuvre consisterait à estimer la masse selon une relation de la forme $M = k \cdot A^b$, où A est l'aire apparente de l'objet, k une constante et b un exposant lié à sa dimensionnalité. Cette approche, séduisante par sa généricité, suppose cependant de connaître à l'avance la forme de l'objet pour calibrer k et b une hypothèse qui ne tient pas pour des objets à géométrie non rigide comme les sachets plastiques, dont l'aire apparente varie fortement selon leur état de déploiement ou de compression. Le développement d'un modèle de classification de forme en amont, capable de distinguer les corps rigides des corps souples avant l'application de la correction appropriée, constituerait une extension naturelle de la classification par élongation déjà mise en œuvre dans ce travail.

Optimisation du modèle de détection. Au-delà de l'estimation de masse, le modèle de détection lui-même pourrait encore être amélioré selon deux axes indépendants. D'une part, sur le plan du compromis précision/latence : l'entraînement d'un modèle de grande taille (XL), suivi d'une distillation de ses connaissances vers un modèle compact compatible avec la cible embarquée, n'a pas été testé faute de temps dans le cadre de cette étude, mais constituerait une alternative à la comparaison directe des variantes n/s/m menée ici. D'autre part, sur le plan de la fidélité géométrique des boîtes englobantes : à l'instar de la perspective évoquée par Grimmer et al., 2025, l'intégration de modèles à boîtes englobantes orientées (*Oriented Bounding Box*, disponibles depuis YOLOv11-OBB et intégrés à YOLO26) permettrait de lever la contrainte actuelle liée aux boîtes axées sur les axes de l'image, qui déforme l'estimation dimensionnelle des objets inclinés par rapport à l'axe optique de la caméra.

Extension du dispositif de terrain. Ces améliorations méthodologiques resteraient enfin limitées sans un élargissement du dispositif de terrain lui-même. Conformément à l'objectif initial du projet, l'augmentation du nombre de caméras déployées sur le bassin versant de Strasbourg permettrait d'étendre le suivi automatisé à l'entièreté du réseau hydrographique, tout en enrichissant la diversité des conditions d'acquisition disponibles pour l'entraînement. L'intégration de conditions pluvieuses, absentes du protocole actuel comme de celui de Grimmer et al., 2025, constituerait également une extension pertinente, la pluie étant à la fois associée à des apports accrus de déchets et à une dégradation spécifique de la qualité d'image (turbulence de surface, gouttes sur l'objectif, variations rapides de luminosité).

6 Conclusion

Face au constat dressé en introduction celui de rivières agissant comme le principal vecteur de transfert des déchets plastiques vers les océans, un phénomène appelé à s'aggraver selon les projections à l'horizon 2050 (Lebreton & Andrady, 2019) ainsi qu'aux lacunes identifiées dans l'état de l'art concernant la surveillance continue, nocturne et quantitative des déchets en rivière, ce travail montre que l'association de caméras in situ, de l'apprentissage profond et de la géométrie projective constitue un outil efficace pour la détection et la quantification des macro-déchets anthropiques en rivière. Il répond aux quatre questions de recherche posées en introduction.

La première portait sur l'amélioration du modèle de détection par rapport à la référence établie par Grimmer et al., 2025. La confrontation des scores obtenus montre un mAP50 inférieur à celui rapporté par cette étude, compris entre 0,300 et 0,786, contre 0,933 à 0,965. Cet écart ne traduit pourtant pas une régression de performance : il reflète un protocole d'évaluation plus strict, retenu pour limiter la fuite de données parmi le volume important de données disponibles. La mise en place d'une validation croisée organisée par site limite en effet la fuite de données, identifiée dans la littérature comme responsable d'une surestimation de plus de 40 points de mAP50. La forte variabilité observée entre sites de test confirme que la performance dépend davantage de la similarité entre site d'entraînement et site de test que du seul volume de données disponible. L'idée de départ, selon laquelle augmenter la quantité de données suffirait à améliorer un modèle, a donc surtout permis de mettre en évidence l'impact déterminant de la diversité spatiale des sites d'acquisition sur la capacité de généralisation. L'élargissement du jeu de données à de nouveaux sites, complété par des approches semi-supervisées exploitant les nombreuses séquences vidéo encore non annotées, constitue à ce titre la piste la plus directe pour consolider ce résultat.

La deuxième hypothèse sugérait l'intérêt d'un modèle unique, détectant les débris de jour comme de nuit, pour limiter l'empreinte du système embarqué. Les résultats obtenus sont encourageants : le modèle jour atteint un mAP50 de 0,877 sur ses propres données, tandis que le modèle mixte, entraîné sur l'ensemble jour et nuit, obtient un mAP50 de 0,836 sur les données nocturnes un score supérieur aux 0,761 obtenus par le modèle nocturne spécialisé sur ces mêmes données. Cette proximité entre les deux régimes, renforcée par la supériorité du modèle mixte sur le modèle spécialisé, suggère en faveur d'un modèle ambivalent plutôt que de deux modèles distincts. Le gain effectif en empreinte mémoire sur la carte mère qu'apporterait un tel modèle unique n'a toutefois pas été mesuré directement sur la cible embarquée elle-même : cette vérification reste la validation manquante pour trancher définitivement la question.

La troisième, portait sur l'apport de données acquises en faible luminosité pour la détection nocturne. Le modèle mixte, entraîné avec ces données, atteint de nuit un niveau de performance proche du niveau diurne, ce qui va dans ce sens. En l'absence d'une comparaison contrôlée avec un modèle entraîné exclusivement sur données diurnes puis testé de nuit, cette amélioration reste néanmoins attribuée plutôt que strictement isolée : une ablation dédiée, opposant directement les deux protocoles d'entraînement, constituerait la validation manquante.

Enfin, la dernière hypothèse portait sur la capacité des caractéristiques géométriques extraites des détections à produire des indicateurs quantitatifs de taille, de volume et de masse suffisamment précis. Les résultats sont ici contrastés. La correction géométrique proposée, qui ajoute à la méthode de Grimmer et al., 2025 une classification morpho-

logique par élongation et des garde-fous physiques, réduit l'erreur moyenne de plusieurs centaines de centimètres à quelques centimètres sur les catégories calibrées, ce qui constitue une amélioration robuste et clairement démontrée. Si la chaîne géométrique enrichie montre sa capacité à corriger les dérives quadratiques de la méthode d'origine en laboratoire, son application in-situ met en lumière la sensibilité de la projection monoculaire aux bruits de détection haute fréquence. La saturation observée sur les pistes réelles montre que la quantification massique automatique requiert désormais l'implémentation d'un filtrage des sauts à l'échelle de la trame et l'exploitation conjointe du score de confiance cumulatif C_{conf} : ce dernier point n'est validé qu'en partie. Un filtrage des sauts de mesure appliqué au niveau de chaque trame, avant l'agrégation par piste, constitue la piste corrective la plus immédiate pour réduire cette saturation.

Au total, ce travail établit une chaîne de traitement complète, de la détection embarquée à l'estimation de masse, et apporte une contribution méthodologique claire sur l'évaluation rigoureuse de la généralisation spatiale et sur la correction des dérives de la régression géométrique de Grimmer et al., 2025. Les deux modèles proposés, jour et mixte, présentent des capacités de détection satisfaisantes en conditions fluviales complexes, tandis que le modèle géométrique corrigé montre une nette réduction de l'erreur d'estimation dimensionnelle par rapport à son origine. L'intégration de l'ensemble de cette chaîne sur cible embarquée ouvre la voie à un déploiement du dispositif à l'échelle du réseau hydrographique de l'Eurométropole de Strasbourg, sous réserve de lever les limites identifiées sur la fiabilité de l'estimation de masse en conditions réelles et sur la validation ciblée du modèle jour/nuit. À terme, un tel dispositif fournirait aux gestionnaires des milieux aquatiques les indicateurs quantitatifs nécessaires au dimensionnement des installations de récupération et à l'organisation des campagnes de collecte, contribuant ainsi, à l'échelle locale, à l'effort plus large de réduction des flux de plastique vers les milieux marins.

Références

Alotaibi, E., & Nassif, N. (2024). Artificial intelligence in environmental monitoring : in-depth analysis. *Discover Artificial Intelligence*. https://doi.org/10.1007/S44163-024-00198-1

Apicella, A., Isgrò, F., & Prevete, R. (2025). Don't push the button ! Exploring data leakage risks in machine learning and transfer learning. *Artificial Intelligence Review*. https://doi.org/10.1007/S10462-025-11326-3

Barnes, D. K., Galgani, F., Thompson, R. C., & Barlaz, M. (2009). Accumulation and fragmentation of plastic debris. *Phil. Trans. R. Soc. B*. https://doi.org/10.1098/rstb.2008.0205

Bianchi, O., & Putro, H. P. (2024). Artificial Intelligence in Environmental Monitoring : Predicting and Managing Climate Change Impacts. *International Transactions on Artificial Intelligence (ITALIC)*. https://doi.org/10.33050/ITALIC.V3I1.650

Chardon, V., Grimmer, G., & Wenger, R. (2025). A simple and low-cost method using active RFID to survey plastic mobility. *Science of The Total Environment (STOTEN)*. https://doi.org/10.1016/j.scitotenv.2025.180336

Das, A., et al. (2026). UAV image dataset... generated using LabelImg and Roboflow. *Data in Brief*. https://pubmed.ncbi.nlm.nih.gov/41675364

Dey, A. (2016). Machine Learning Algorithms : A Review. *International Journal of Computer Science and Information Technology*. https://ijcsit.com/docs/Volume%207/vol7issue3/ijcsit2016070332.pdf

Galgani, L., Beiras, R., Galgani, F., & Panti, C. (2019). Impacts of marine litter. *Frontiers in Marine Science*. https://www.frontiersin.org/journals/marine-science/articles/10.3389/fmars.2019.00208/full

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., & Zhu, X. X. (2022). A Survey of Uncertainty in Deep Neural Networks. https://arxiv.org/abs/2107.03342

Gou, J., et al. (2021). Knowledge Distillation : A Survey. *International Journal of Computer Vision (IJCV)*. https://doi.org/10.1007/S11263-021-01453-Z

Grimmer, G., et al. (2025). A geometric and deep learning pipeline for monitoring debris in urban rivers. *arXiv preprint arXiv :2510.23798*. https://doi.org/10.48550/arXiv.2510.23798

Hoiem, D., Gupta, T., Li, Z., & Shlapentokh-Rothman, M. (2021, 18–24 Jul). Learning Curves for Analysis of Deep Networks. In M. Meila & T. Zhang (éd.), *Proceedings of the 38th International Conference on Machine Learning* (p. 4287-4296, T. 139). PMLR. https://proceedings.mlr.press/v139/hoiem21a.html

Hossain, M., & Engelhardt, I. (2025). Global plastic footprint... *Energy Ecology and Environment*. https://doi.org/10.1007/S40974-025-00383-7

Howard, A. G., et al. (2017). MobileNets : Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv preprint arXiv :1704.04861*. https://doi.org/10.48550/arXiv.1704.04861

Hussain, M. (2023). YOLO-v1 to YOLO-v8... *Machines*. https : / / doi . org / 10 . 3390 / MACHINES11070677

Ilani, M. A., & Banadaki, Y. M. (2025). LabelImg : CNN-Based Surface Defect Detection. *arXiv preprint arXiv :2509.05813*. https://doi.org/10.48550/ARXIV.2509.05813

Islam, K., et al. (2024). Diffusemix : Label-Preserving Data Augmentation with Diffusion Models. *CVPR*. https://doi.org/10.1109/CVPR52733.2024.02608

Jocher, G., & Qiu, J. (2026). *Ultralytics YOLO26* (Version 26.0.0). https://github.com/ultralytics/ultralytics

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning : Trends, perspectives, and prospects. *Science*. https://doi.org/10.1126/SCIENCE.AAA8415

Kumar, R. (2017). Machine Learning—Basics. In *Apress eBooks*. https://doi.org/10.1007/978-1-4842-3069-5_4

Kunwar, S. D., & Owabumoye, B. R. (2025). Plastic Waste Detection : Insights from the WaDaBa Dataset. *IDWM*. https://doi.org/10.53623/IDWM.V5I1.580

Lavender, S. (2022). Detection of Waste Plastics... Application of Copernicus EO data. *Preprints*. https://doi.org/10.20944/PREPRINTS202207.0410.V2

Lebreton, L., et al. (2017). River plastic emissions to the world's oceans. *Nature Communications*. https://doi.org/10.1038/NCOMMS15611

Lebreton, L., & Andrady, A. L. (2019). Future scenarios of global plastic waste generation and disposal. *Palgrave Communications*. https://doi.org/10.1057/S41599-018-0212-7

Lebreton, L., de Vries, R., Pham, Y., Wolter, H., van Vulpen, M., Puskic, P., Sainte-Rose, B., Royer, S.-J., & Egger, M. (2024). Seven years into the North Pacific garbage patch : legacy plastic fragments rising disproportionally faster than larger floating objects. *Environmental Research Letters*, *19*(12), 124054. https://doi.org/10.1088/1748-9326/ad78ed

LeCun, Y. (2016). L'apprentissage profond, une révolution en intelligence artificielle. *La lettre du Collège de France*. https://doi.org/10.4000/LETTRE-CDF.3227

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444. https://doi.org/10.1038/nature14539

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278-2324. https://doi.org/10.1109/5.726791

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature Pyramid Networks for Object Detection. https://arxiv.org/abs/1612.03144

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD : Single Shot MultiBox Detector. In *Computer Vision – ECCV 2016* (p. 21-37). Springer International Publishing. https://doi.org/10.1007/978-3-319-46448-0_2

Lusher, A. L., Welden, N. A., Sobral, P., & Cole, M. (2017). Sampling, isolating and identifying microplastics ingested by fish and invertebrates. *Analytical Methods*, *9*(9), 1346-1360. https://doi.org/10.1039/c6ay02415g

Luz, T. M. d., et al. (2025). Tracing Plastic Origins : Role of Ocean Currents and Shipping. *Earth Systems and Environment*. https://doi.org/10.1007/S41748-025-00600-7

Martínez-Hernández, U., et al. (2024). Low-Cost Recognition of Plastic Waste Using Deep Learning and a Multi-Spectral Near-Infrared Sensor. *Sensors*. https://pubmed.ncbi.nlm.nih.gov/38732925

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., & Terzopoulos, D. (2022). Image Segmentation Using Deep Learning : A Survey. *IEEE TPAMI*. https://doi.org/10.1109/TPAMI.2021.3059968

Moore, C. J., et al. (2011). Plastic debris flowing from two urban rivers to coastal waters. *Revista de Gestão Costeira*. https://doi.org/10.5894/RGCI194

Nakayama, T., & Osako, M. (2023). The flux and fate of plastic in the world's major rivers. *Global and Planetary Change*.

Naqa, I. E., & Murphy, M. J. (2015). What Is Machine Learning ? In *Machine Learning in Radiation Oncology*. https://doi.org/10.1007/978-3-319-18305-3_1

Nguyen Tien, S., Le Tien, T., Ly-Thanh, H., & Ly Tu, N. (2026). ONNX-based Architectures for Post-Training Quantization Face Detection on Edge Devices. https://doi.org/10.15598/aeee.v24ix.251003

Noria Innovateurs durables. (s. d.). Caméra IA [Consulté en 2026]. https://www.noria.earth/noria-home/ai-cam/

Obscape. (s. d.). Pressure Wave Gauge [Consulté en 2026]. https://site.obscape.com/products/wave-monitoring/pressure-wave-gauge/

Pan, S. J., & Yang, Q. (2009). A Survey on Transfer Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/TKDE.2009.191

Pereira, F., Mitchell, T. M., & Botvinick, M. (2009). Machine learning classifiers and fMRI : a tutorial overview. *NeuroImage*. https://pubmed.ncbi.nlm.nih.gov/19070668

Perez-Torrents, J., & Minvielle, É. (s. d.). Comment les usages des outils d'Intelligence Artificielle augmentent-ils la capacité d'agir des radiologues ?, 185-203. https://doi.org/10.54695/jdds.041.3.0185

Ramos, E. O. B., et al. (2024). Application of ML in Plastic Waste Detection : A Systematic Review. *Processes*. https://doi.org/10.3390/PR12081632

Redmon, J., & Farhadi, A. (2018). YOLOv3 : An Incremental Improvement. *arXiv preprint arXiv :1804.02767*.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once : Unified, Real-Time Object Detection. https://arxiv.org/abs/1506.02640

Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster R-CNN : Towards Real-Time Object Detection with Region Proposal Networks. https://arxiv.org/abs/1506.01497

Salinas, M., Rosas, J., Iborra, J., Manero, H., & Pascual, E. (1997). Comparison of manual and automated cell counts in EDTA preserved synovial fluids. Storage has little influence on the results. *Annals of the Rheumatic Diseases*. https://doi.org/10.1136/ard.56.10.622

Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge Computing : Vision and Challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/JIOT.2016.2579198

Shobaki, W. A., & Milanova, M. (2025). A Comparative Study of YOLO, SSD, Faster R-CNN. *Sci*. https://doi.org/10.3390/SCI7020047

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*. https://doi.org/10.1186/S40537-019-0197-0

Shukhratov, I., Pimenov, A., Stepanov, A., Mikhailova, N., Baldycheva, A., & Somov, A. (2024). Optical detection of plastic waste through computer vision. *Intelligent Systems with Applications*, *22*, 200341. https://doi.org/https://doi.org/10.1016/j.iswa.2024.200341

Sio, G. A., et al. (2022). Plastic Waste Detection on Rivers Using YOLOv5 Algorithm. *ICCCNT 2022*. https://doi.org/10.1109/ICCCNT54827.2022.9984439

Surfrider Foundation Europe. (2025). *Livre Blanc : L'intelligence artificielle pour quantifier les déchets en rivière - Projet Plastic Origins* (rapp. tech.). Surfrider Foundation Europe. https://www.surfrider.fr/wp-content/uploads/2025/05/Livre_Blanc_IA_quantifier_dechets_riviere-PlasticOrigins_SurfriderFoundationEurope.pdf

Tamin, O., et al. (2022). Machine Learning for Plastic Waste Detection : State-of-the-art, Challenges. *CIEES 2022*. https://doi.org/10.1109/CIEES55704.2022.9990703

Tariq, M. F., & Javed, M. A. (2025). Small Object Detection with YOLO : A Performance Analysis Across Model Versions and Hardware. https://arxiv.org/abs/2504.09900

Thompson, R. C., et al. (2004). Lost at Sea : Where Is All the Plastic ? *Science*. https://pubmed.ncbi.nlm.nih.gov/15131299

Tramoy, R., Gasperi, J., Colasse, L., & Tassin, B. (2020). Transfer dynamic of macroplastics in estuaries — New insights from the Seine estuary : Part 1. Long term dynamic based on date-prints on stranded debris. *Marine Pollution Bulletin*, *152*, 110894. https://doi.org/10.1016/j.marpolbul.2020.110894

Tramoy, R., Gasperi, J., Nguyen, L., & Tassin, B. (2020). Transfer dynamic of macroplastics in estuaries — New insights from the Seine estuary : Part 2. Short-term dynamics based on predicting stranded macroplastic objects. *Marine Pollution Bulletin*, *152*, 110901. https://doi.org/10.1016/j.marpolbul.2020.110901

Ultralytics. (2026). *Ultralytics YOLO Documentation* [Consulté le xx/xx/2026]. https://docs.ultralytics.com/

van Emmerik, T., & Schwarz, A. (2020). Plastic debris in rivers. *WIREs Water*. https://doi.org/10.1002/WAT2.1398

van Emmerik, T., Kieu-Le, T.-C., Loozen, M., van Oeveren, K., Strady, E., Bui, X.-T., Egger, M., Gasperi, J., Lebreton, L., Nguyen, P.-D., Schwarz, A., Slat, B., & Tassin, B. (2018). A Methodology to Characterize Riverine Macroplastic Emission Into the Ocean. *Frontiers in Marine Science*, *Volume 5 - 2018*. https://doi.org/10.3389/fmars.2018.00372

Winterstetter, A., Grodent, M., Kini, V., Ragaert, K., & Vrancken, K. C. (2021). A Review of Technological Solutions to Prevent or Reduce Marine Plastic Litter in Developing Countries. *Sustainability*, *13*(9). https://doi.org/10.3390/su13094894

Wu, P. C. (2026). Evaluating Open-Vocabulary Object Detection in Low-quality Image. *arXiv preprint*.

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., & Loy, C. C. (2022). Domain Generalization : A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1-20. https://doi.org/10.1109/tpami.2022.3195549

7 Annexe

7.1 Déclaration d'utilisation de l'intelligence artificielle

Dans une démarche de transparence, l'intelligence artificielle générative a été utilisée ponctuellement comme outil d'assistance lors de la réalisation de ce mémoire. Cet usage a porté sur l'amélioration de la rédaction, en permettant de mettre en avant les répétitions de texte et d'idées afin de les éliminer, de structurer les idées lors de l'élaboration du plan, d'aider à la bibliographie en proposant un plan d'action pour la recherche dans la littérature, ainsi que de corriger l'orthographe et la grammaire et de vérifier la cohérence terminologique, en s'assurant que les mêmes termes techniques soient employés de façon uniforme dans l'ensemble du document.

Elle a également été employée pour l'optimisation des scripts Python, la création des scripts les plus complexes, notamment ceux liés au bilan de masse, le débogage des erreurs de compilation LaTeX liées à la syntaxe et à la structure des tableaux, ainsi que l'aide à la mise en forme des figures et des tableaux, en particulier pour les couleurs, l'alignement et la largeur, et à la génération de portions de code pour l'analyse de données et les visualisations.

Ces outils n'ayant joué qu'un rôle de support technique, j'ai personnellement pensé, vérifié et validé l'ensemble des protocoles expérimentaux, des données collectées, des analyses et des conclusions. J'assume ainsi l'entière responsabilité du contenu de ce travail.

7.2 Données dans la littérature

Dataset RGB	Images	Prediction	labels	formats	classes
Amsterdam	9		9	txt	1
Chine	1500		1500	VOC-pascal	?
Chine	1500		1500	VOC-pascal	?
Croatie	5490		5490	COCO-YOLO	1
Croatie	5299		5299	COCO-YOLO	1
Ghana	15		/	/	/
Groningen	64		63	YOLO	1
Jakarta	526		526	CSV	3
Japon	7356		102	Json	02- 05- 07
Malaisie	422		422	YOLO	?
Oostpoort	563		562	Yolo	?
Pakistan	12801		12801	VOC-pascal	?
Debrit-Dataset	519		/	/	/
Trash-Dataset	2400	Mauvaise	2400	YOLO	8
Espagne	182		/	/	/
GreenVilage	1501		1501	YOLO	1
USA	124		/	/	/
TUD-GV	77 dossier		/	/	/
TUD-HCMC	29 dossier		/	/	/

7.3 Glossaire

Glossaire

AP (Average Precision) Aire sous la courbe précision-rappel pour une classe donnée, en faisant varier le seuil de confiance des détections.

BoT-SORT Algorithme de suivi multi-objets standard, dérivé de ByteTrack, ajoutant une compensation des mouvements de caméra (GMC) et une réidentification par apparence (Re-ID).

ByteTrack Algorithme de suivi multi-objets standard fondant l'association sur l'IoU, exploitant à la fois les détections à fort et à faible score de confiance.

C_conf (score de confiance cumulatif) Indice compris entre 0 et 1 associé à chaque objet, produit de six facteurs d'atténuation ($\overline{P}_{\text{YOLO}}$, ω_{track}, ω_{calib}, ω_{shape}, ω_{dyn}, ω_{phys}) rendant compte de la fiabilité de chaque étape du pipeline d'estimation de masse.

CNN (Convolutional Neural Network) Réseau de neurones convolutif ; architecture fondatrice de la vision par ordinateur moderne, reposant sur des couches de convolution et de pooling.

Data leakage (fuite de données) Contamination du jeu de test par des informations issues du jeu d'entraînement (même site, même séquence vidéo), conduisant à une surestimation artificielle des performances.

Domain shift / Domain Generalization Sensibilité d'un modèle entraîné sur un domaine donné (une rivière) à des conditions d'acquisition différentes lors du déploiement ; la Domain Generalization vise à développer des modèles robustes à ce changement sans ré-entraînement.

Edge AI Exécution de modèles d'intelligence artificielle directement sur un dispositif embarqué, sans recours à une infrastructure de calcul distante.

E_s (ratio d'élongation) Rapport entre la moyenne du quart supérieur de la longueur projetée et la moyenne du quart inférieur de la hauteur projetée d'une piste ; sert à classer l'objet en bouteille/canette ($E_s \geq 2{,}2$) ou en bidon ($E_s < 2{,}2$).

FOV (Field Of View) Champ de vision horizontal ou vertical de la caméra, dérivé de la focale et des dimensions physiques du capteur.

F1-score Moyenne harmonique de la précision et du rappel, synthétisant leur compromis.

Gap filtering (filtrage des sauts) Traitement statistique appliqué à une piste de suivi pour éliminer les valeurs aberrantes de dimensions projetées, selon un seuil dynamique $\Delta_{\max} = \max(1{,}5\,\sigma,\ 2{,}5\text{ cm})$.

GFLOPs Milliards d'opérations en virgule flottante par seconde ; mesure du coût de calcul d'un modèle.

imgsz Taille (en pixels) de l'image d'entrée du réseau de détection.

IoU (Intersection over Union) Rapport entre l'aire d'intersection et l'aire d'union de deux boîtes englobantes ; mesure le recouvrement spatial entre une prédiction et une annotation.

K-fold (validation croisée) Méthode de validation qui divise le jeu de données en K sous-ensembles, chacun servant tour à tour de jeu de validation, adaptée ici par site de caméra plutôt que par tirage aléatoire.

mAP (mean Average Precision) Moyenne des AP sur l'ensemble des classes ; mAP50 (seuil IoU = 0,5) et mAP50-95 (moyenne sur seuils IoU de 0,5 à 0,95).

MOT (Multi-Object Tracking) Suivi multi-objets ; association des détections successives d'un même objet sous un identifiant de piste (Track ID) unique.

NCNN Bibliothèque d'inférence légère développée par Tencent, optimisée pour les processeurs ARM embarqués (instructions ARM NEON).

NMS (Non-Maximum Suppression) Étape de filtrage des boîtes englobantes redondantes ; supprimée dans l'architecture YOLO26 au profit d'un apprentissage de bout en bout.

ONNX Format d'échange universel pour modèles de deep learning, facilitant l'interopérabilité entre outils d'inférence.

OpenVINO Moteur d'inférence développé par Intel, optimisé pour les architectures matérielles x86_64 et les accélérateurs NPU.

Pinhole (modèle sténopé) Modèle de caméra simplifié utilisé pour la reprojection géométrique des boîtes englobantes vers des dimensions réelles, fondé sur l'intersection de rayons optiques avec le plan d'eau.

Precision (précision, P) Proportion de détections effectivement correctes parmi l'ensemble des détections produites : $P = VP/(VP + FP)$.

Recall (rappel, R) Proportion d'objets réellement présents retrouvés par le modèle : $R = VP/(VP + FN)$.

ρ_{mat} (densité apparente) Densité matière équivalente de l'enveloppe plastique d'un débris creux, retenue dans la plage [0,024 ; 0,0356] g/cm³.

STAL (Small-Target-Aware Label Assignment) Mécanisme intégré à YOLO26 garantissant qu'un objet de petite taille reçoive toujours au moins une assignation positive lors de l'entraînement.

TOPS (Tera Operations Per Second) Unité de mesure de la puissance de calcul d'un accélérateur matériel (ex. Hailo-8L).

Track ID Identifiant unique attribué à un objet suivi à travers les trames successives d'une séquence vidéo.

TYP1 / TYP2 / TYP3 Les trois niveaux de la typologie de classification des débris développée dans ce travail : nature du débris (anthropique/naturel/non-débris), composition (matériau), puis structure fine (rigide/flexible, éléments constitutifs).

YOLO (You Only Look Once) Famille d'architectures de détection d'objets en une seule étape (one-stage), privilégiant le compromis vitesse/précision ; YOLO26 est la version utilisée dans cette étude (Ultralytics, 2026).

Faculté de géographie et d'aménagement
Université de Strasbourg

Master Observation de la Terre et Géomatique

2025–2026

RÉSUMÉ :
La pollution plastique des cours d'eau est un vecteur majeur de contamination des milieux marins. Ce travail propose une chaîne de traitement complète pour la détection et la quantification massique automatisées de macro-déchets plastiques flottants, déployée sur des caméras fixes obliques sur la Bruche et le Steingiessen (Eurométropole de Strasbourg).

Un jeu de données en images couleur et infrarouge (jour/nuit) a été constitué sur quatre sites et combiné aux données de Grimmer et al., 2025. Plusieurs architectures YOLO ont été comparées et évaluées par validation croisée par site, limitant les fuites de données et mesurant la généralisation spatiale. Le modèle retenu a été déployé sur Raspberry Pi 5 au format NCNN (FP16).

Un tracker géométrique dédié assure le suivi des trajectoires en rivière. Les dimensions réelles sont estimées par projection géométrique monoculaire (Grimmer et al., 2025), complétée ici par une classification par élongation et des garde-fous réduisant l'erreur moyenne de plusieurs centaines de centimètres à quelques centimètres. L'estimation de masse reste limitée par des saturations observées sur données réelles, piste de travail prioritaire.

MOTS-CLÉS :
IA embarqué, Détection d'objet, Apprentissage profond, YOLOv8/YOLOv26, Raspberry Pi 5, Surveillance plastique en cours d'eau

ABSTRACT :
Plastic pollution in waterways is a major vector of contamination in marine environments. This work proposes a complete processing pipeline for the automated detection and mass quantification of floating plastic macro-debris, deployed using fixed oblique cameras on the Bruche and the Steingiessen (Eurometropolis of Strasbourg).

A dataset of color and infrared images (day/night) was compiled across four sites and combined with data from Grimmer et al., 2025. Several YOLO architectures were compared and evaluated through cross-validation by site, limiting data leakage and measuring spatial generalization. The selected model was deployed on a Raspberry Pi 5 in NCNN (FP16) format.

A dedicated geometric tracker ensures trajectory tracking in river conditions. Real dimensions are estimated through monocular geometric projection (Grimmer et al., 2025), complemented here by an elongation-based classification and safeguards reducing the mean error from several hundred centimeters to a few centimeters. The resulting mass estimation remains limited by saturation phenomena observed on real data, identified as a priority direction for future work.

KEYWORDS :
Edge AI, object detection, Deep learning, YOLOv8/YOLOv26, Raspberry Pi 5, river debris monitoring

Laboratoire image, ville et environnement | LIVE
Université de Strasbourg & cnrs & ENGEES
Strasbourg.eu eurométropole
cnrs