



UE PRAT

Estimation du flot optique par apprentissage non supervisé

Rapport finale

Auteur :

BOUMERDAS Wassila
28617059

Encadré par :

Dominique BÉRÉZIAT

Problématique :

De nos jours, même si on ignore la signification exacte du terme flot optique son estimation joue un rôle cruciale dans de nombreuses applications, comme l'estimation du mouvement et la compression vidéo, la détection et le suivi d'objets, l'odométrie visuelle. Ce concept de flot optique a été proposé par le psychologue James Jerome Gibson en 1950 dans une étude sur la vision humaine. En 1980, Hornet et Schunck ont proposé une méthode d'estimation du flot optique basée sur la régularisation. Ce premier travail a été suivi d'un grand nombre de contributions qui ont proposé différentes méthodes alternatives afin de pallier aux différents aspects de cet ancien enjeu qui restent largement non résolus à ce jour.

Les premières approches traitant le problème du flux optique vise à minimiser pour une paire d'images une fonction de perte, en intégrant des hypothèses fortes sur la luminosité et la fluidité spatiale du pixel. Par la suite, des formulations de champ aléatoire de Markov (MRF) ont été exploitées pour pallier aux difficultés rencontrées avec les approches traditionnels. Mais avec le succès remarquable de CNN dans la compréhension des images, les méthodes d'apprentissage approfondi proposées ont été appliqués avec succès au problème du flux optique, telles que FlowNetS/FlowNetC, FlowNet2.0, PWC-Net, etc. L'exploitation de ses approches à révolutionner le domaine du flot optique mais sans pour autant qu'elle soit la solution la plus optimale et sans défauts. En effet, l'apprentissage supervisée de tels modèles nécessite de grands ensembles de données étiquetées ce qui reste un défi vu la difficulté de l'obtention de ses dernières. Ainsi, les approches existantes s'appuient principalement sur des données synthétiques ce qui induit une inadéquation entre les données d'apprentissage et les données de test. Pour la difficulté d'obtenir des étiquettes de correspondance denses et précises, l'apprentissage non supervisé du flux optique a attiré de plus en plus d'attention, de nombreux travaux récents ont proposé d'apprendre le flux optique de manière non supervisée, dans laquelle la vérité de base n'est pas nécessaire ce qui pourrait vraisemblablement produire des résultats de meilleure qualité.

L'hypothèse de base partagée par les techniques de flux optique non supervisées est que l'apparence d'un objet ne change pas lorsqu'il se déplace, ce qui permet d'entraîner ces modèles en utilisant la vidéo sans étiquette comme suit : Le modèle est utilisé pour estimer un champ de flux entre deux images I_1 et I_2 . Ce champ de flux est utilisé pour modifier l'une des images afin qu'elle corresponde à l'autre, puis les poids du modèle sont mis à jour de manière à minimiser la différence entre ces deux images - et à le permettre une certaine forme de régularisation[8]. Bien que toutes les méthodes de flux optique non supervisées partagent cette même idée de base, leurs détails varient grandement, mais encore, leurs performances restent encore insatisfaisantes et souvent très en retard par rapport à leurs homologues supervisées, principalement en raison d'un lissage excessif des limites de mouvement et de l'occlusion. [17] Dans ce sens les chercheurs ne cessent de tester, comparer, améliorer et intégrer systématiquement les différentes approches et leurs paramètres afin d'approfondir leurs compréhension et de fournir de nouvelles méthodes plus robustes dans l'estimation du flux optiques.

État de l'art

Introduction

Dans la première partie de ce document, nous nous intéressons à la définition des différents concepts portant sur la définition, caractérisation et le calcul du flot optique et les différents concepts liés à ce dernier en se basant sur deux sources de références dans le domaine du flot optique. On décrira ainsi les approches de calcul du flot optique existante, de manière à effectuer des premiers choix.

Définition

Le flux optique est la distribution des vitesses apparentes de mouvement mesuré à partir des variations de la luminosité dans une image. Il peut résulter du mouvement relatif des objets et de l'observateur [5].

Hypothèse et notations

Hypothèse : les séquences d'image sont (localement) invariantes en luminosité, ie, il y a une conservation des valeurs d'image le long des trajectoires.

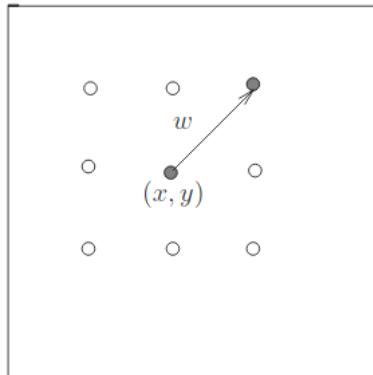


FIGURE 1 – Transport de la luminosité

Notation :

Selon l'hypothèse de constance de la luminosité, l'intensité du pixel reste la même malgré de petits changements de position et de période. Plus précisément,

$$I(x, y, t) = I(x + \delta_x, y + \delta_y, t + \delta_t), \forall (x, y) \in \Omega$$

Avec :

- Ω : domaine de l'image (région bornée de R)
- $w = (\delta_x, \delta_y)$: déplacement du point (x, y) au temps t

Où $(\delta_x, \delta_y, \delta_t)$ est le petit changement du mouvement. $I(x + \delta_x, y + \delta_y, t + \delta_t)$ peut être exprimé par une expansion en série de Taylor :

$$I(x + \delta_x, y + \delta_y, t + \delta_t) = I(x, y, t) + \delta_x \frac{\partial I}{\partial x}(x, y, t) + \delta_y \frac{\partial I}{\partial y}(x, y, t) + \delta_t \frac{\partial I}{\partial t}(x, y, t)$$

Par conséquent,

$$\begin{aligned}\delta_x \frac{\partial I}{\partial x} + \delta_y \frac{\partial I}{\partial y} + \frac{\partial I}{\partial t} &= 0 \\ \frac{\partial I}{\partial x} U_x + \frac{\partial I}{\partial y} U_y + \frac{\partial I}{\partial t} &= 0\end{aligned}$$

avec $U_x = \frac{\delta_x}{\delta t}$, $U_y = \frac{\delta_y}{\delta t}$ et $w = (\frac{\delta x}{\delta t}, \frac{\text{deltay}}{\delta t})$ et $\nabla I = (\frac{\text{partialI}}{\partial x}, \frac{\text{partialI}}{\partial y})$.

on aura alors l'equation de la contrainte du flot optique (Optical Flow Constraint Equation) :

$$\omega \cdot \nabla I + I_t = 0 \quad (1)$$

L'hypothèse de conservation de la luminosité est forte car elle suppose que les variations temporelles de luminance sont uniquement dues au mouvement. On a par ailleurs gardé simplement les termes du premier ordre.

Cette contrainte est insuffisante pour déterminer le flot complet ω car le problème est mal posé : on dispose d'une équation linéaire pour deux inconnues.

Cela signifie que l'on est simplement en mesure d'évaluer ω_{\perp} , la vitesse normale dirigée selon le gradient local. Concrètement, on est en présence de trop de solutions possibles pour évaluer la solution réelle, il faut donc particulariser la solution, ajouter une contrainte de régularisation notamment.

On peut évaluer la vitesse normale de la manière suivante : Suite à (1), on a $\vec{\omega} = \frac{I_t}{|\nabla I|}$.

On définit par ailleurs le vecteur normal $\vec{n} = \frac{\nabla I}{|\nabla I|}$.

La projection de la vitesse sur la normale est alors $s = \omega^T n$.

La vitesse normale vaut $\omega_{\perp} = (\vec{\omega} \cdot \vec{n}) n = -\frac{I_t}{|\nabla I|} \cdot \frac{\nabla I}{|\nabla I|} \cdot \frac{\nabla I}{|\nabla I|} = -\frac{I_t \nabla I}{|\nabla I|^2}$

Le flot optique est exactement équivalent au mouvement image si les hypothèses suivantes sont respectées :

- éclairage uniforme–surface à réflexivité Lambertienne
- mouvement de translation pure,
- parallèle au plan de l'image

Ces conditions ne sont bien sûr jamais rigoureusement respectées dans le cas d'images réelles. On peut toutefois les supposer vérifiées localement. Une fois que le flux optique est calculé, il peut être utilisé pour apprendre la représentation au niveau vidéo pour la reconnaissance des gestes.

Les approches d'estimation du flot optique

Le mouvement entre un objet et un spectateur provoque un mouvement apparent des motifs de luminosité dans l'image. Les techniques de flux optique tentent d'inverser cette relation pour récupérer une estimation du mouvement. Les méthodes classiques déduisent un flux optique pour une paire d'images en minimisant une fonction de perte qui mesure la cohérence et la fluidité photométriques[5]. Les approches récentes considèrent l'estimation du flux optique comme un problème d'apprentissage dans lequel un modèle basé sur CNN régresse d'une paire d'images à un champ de flux [3],[6]. Certains modèles intègrent des idées de méthodes antérieures, telles que les volumes de coûts et le coarse-to-fine (Estimation du flux progressif des pyramides gaussiennes de niveau grossier à fin des deux images)[16][18]. Comme de telles approches sont de faible ampleur, les méthodes supervisées ont principalement reposé sur des données synthétiques pour la formation, et souvent pour l'évaluation. La synthèse de "bonnes" données d'entraînement (telles que les modèles appris se généralisent aux images réelles) est en soi un problème de recherche difficile, nécessitant un examen minutieux du contenu de la scène, du mouvement de la caméra, de la distorsion de l'objectif et de la dégradation du capteur. Les approches non supervisées contournent le besoin d'étiquettes en optimisant la cohérence pho-tométrique avec une certaine régularisation[8], similaire aux méthodes classiques basées sur l'optimisation mentionnées ci-dessus. Alors que les méthodes traditionnelles résolvent un problème d'optimisation pour chaque paire d'images, l'apprentissage non supervisé optimise conjointement un objectif sur toutes les paires d'un ensemble de données et apprend une fonction qui régresse un champ de flux à partir des images. Cette approche présente deux avantages :

- l'inférence est rapide car l'optimisation n'est effectuée que pendant la formation,
- en optimisant conjointement l'ensemble du train, les informations sont partagées entre les paires d'images, ce qui peut potentiellement améliorer les performances.

Quelques exemples d'approches détaillées :

Nous détaillerons dans ce qui suit des approches que nous estimons intéressantes dans la compréhension dans la suite de notre travail :

Approche supervisée PWCNET :

Principe générale de l'approche PWC-NET

Le PWC-Net [16] a été conçu selon des principes simples et bien établis : traitement pyramidal, warping et utilisation d'un volume de coûts. Fondée dans une pyramide de caractéristiques apprises, PWC-Net utilise l'estimation du flux optique pour déformer (warp) les caractéristiques CNN de la deuxième image. Elle utilise ensuite les caractéristiques déformées (warped) et les caractéristiques de la première image pour construire un volume de coût, qui est traité par une CNN pour estimer le flux optique. Le PWCNet est 17 fois plus petit et plus facile à former que FlowNet2. En effet, il apporte des améliorations significatives en termes de taille et de précision des modèles par rapport aux modèles CNN existants (flownet et flownet2 [3][6]) pour les flux optiques.

Architecture :

La figure 2 résume les principales composantes du PWC-Net. Tout d'abord, le modèle est composé d'un extracteur pyramide de caractéristiques. Ensuite, on a une couche de déformation (warping) qui est utilisée comme une couche dans le réseau pour estimer les grands mouvements. Le réseau dispose également d'une couche pour construire le volume de coût, qui est ensuite traité par les couches CNN pour estimer le flux. Les couches de déformation (warping) et de volume de coût n'ont pas de paramètres apprentissables et réduisent la taille du modèle. Enfin, PWC-Net utilise un réseau de contexte pour exploiter les informations contextuelles afin d'affiner le flux optique. Nous expliquerons les idées principales de chaque composant, notamment l'extracteur de caractéristiques de la pyramide, l'estimateur de flux optique et les réseaux contextuels.

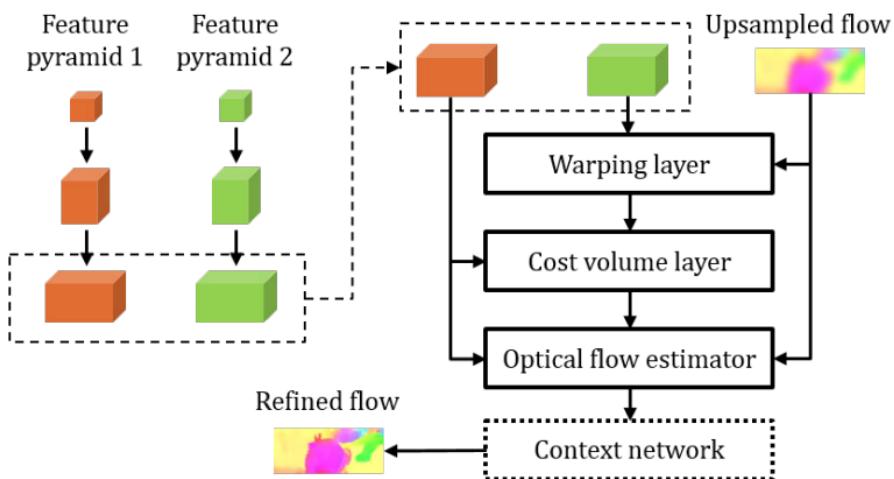


FIGURE 2 – Architecture du modèle PWCNET

Extracteur pyramide de caractéristique

PWC-Net utilise un CNN à deux couches pour extraire les caractéristiques à chaque niveau de la pyramide. En effet, l'idée est de générer des pyramides de l-niveau de représentations des caractéristiques des images, le niveau inférieur (zéro) étant les images d'entrée. Pour générer la représentation des caractéristiques au niveau de la l-ième couche, des couches de filtres convolutifs sont utilisées pour échantillonner les caractéristiques au niveau de la l-1 ème pyramide, par un facteur de 2.

Wrapping couche :

Au niveau 1 et grâce à la couche warping, un warp est effectué sur la 2ème image en utilisant le flux suréchantillonné x2 du niveau l+1 afin d'obtenir la première image $\sim I_1$ *Cost volume couche Les caractéristiques sont utilisées pour construire un volume de coûts qui stocke les coûts d'association d'un pixel avec ses pixels correspondants à l'image suivante. *Optical flow estimator Il s'agit d'un CNN multicouche. Ses entrées sont le volume de coût, les caractéristiques de la première image et le flux optique sur-échantillonné et sa sortie est le flux w_l au l ème niveau. Les estimateurs aux différents niveaux ont leurs propres paramètres au lieu de partager les mêmes paramètres. Ce processus d'estimation est répété jusqu'au niveau souhaité. Les entrées de chaque couche convolutive sont la sortie et l'entrée de la couche précédente. PWC-Net utilise un CNN à cinq couches dans l'estimateur de flux optique à chaque niveau.

*Context network un réseau de contexte sert ici à exploiter les informations contextuelles afin d'affiner le flux optique. C'est un CNN de type feed-forward et sa conception est constitué de 7 couches convolutives. Le noyau spatial pour chaque couche convulsive est de 3×3 . Ces couches ont des constantes de dilatation différentes. De bas en haut, les constantes de dilatation sont 1, 2, 4, 8, 16, 1 et 1.

Conclusion

PWC-Net utilise un traitement pyramidal pour augmenter la résolution de manière globale à fine et utilise le warping des caractéristiques, la construction du volume des coûts pour estimer le flux optique à chaque niveau. Sur la base de ces principes, il a atteint des performances de pointe avec un modèle de taille compacte. basée sur ses performances et ses qualités, cette approche est utilisé comme architecture de base dans plusieurs Approches non-supervisé.

Approche DDflow :

Self-supervision et Data distillation :

Avant d'introduire cet approche nous expliquons d'abord ce que c'est la self-supervision et la distillation des connaissances.

En effet,l'idée de la self-supervision dans les flux optiques non supervisés est de générer des étiquettes de flux optiques en appliquant le modèle appris sur une paire d'images, puis de modifier les images pour rendre l'estimation du flux plus difficile et d'entraîner le modèle à récupérer le flux estimé à l'origine.

Or que nous définissons la distillation des connaissance comme étant processus de transfert des connaissances d'un grand modèle à un plus petit sans perte de validité. Comme les modèles plus petits sont moins coûteux à évaluer, ils peuvent être déployés sur du matériel moins puissant. proposons de mieux expliquer ce que s'est la distillation des données.

Principe générale de l'approche DDFLOW [10]

Cette approche basée sur le principe de la distillation de données vise à optimiser deux modèles, un modèle d'enseignant et un modèle d'étudiant (comme le montre la figure 1). Pour se faire, l'idée est d'entraîner le modèle de l'enseignant à estimer le flux optique pour les pixels non occlus (par exemple, (x_1, y_1) dans I_1). Ensuite, l'occlusion de flux est provoquée en recordant les images originales (le pixel (x_1, y_1) devient alors occlus dans \tilde{I}_1). Les prédictions du modèle enseignant sont utilisées comme annotations pour guider directement le réseau d'étudiants vers l'apprentissage du flux optique. Les deux réseaux partagent la même architecture et sont formés de bout en bout avec de simples fonctions de perte. Le réseau d'étudiants est utilisé pour produire un flux optique au moment du test et fonctionne en temps réel Afin de mieux comprendre le principe générale de l'approche DDFlow, supposons qu'on a deux images I_1 et I_2 , avec (x_2, y_2) pixel dans I_2 qui correspond au pixel (x_1, y_1) dans I_1 . Si on prend (x_1, y_1) pixel non occlus dans I_1 , on peut juste utiliser la perte photo-métrique classique afin d'estimer son flux optique dans le modèle d'enseignant. Ensuite au niveau du modèle étudiant, si nous recordons nos images on prenons des patchs \tilde{I}_1 and \tilde{I}_2 , le pixel (x_1, y_1) devient occulté, puis que il n'a plus sont correspondant dans le patch \tilde{I}_2 .Pour y remédier, l'idée est d'utiliser la prédiction de l'enseignant comme annotation pour superviser le modèle d'élève dans son apprentissage pour ce pixel occlus. C'est l'intuition clé qui se trouve derrière le DDFlow.

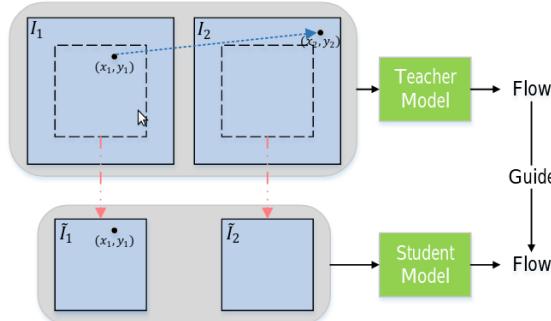


FIGURE 3 – Illustration du principe générale basée sur la distillation des connaissance de DDFlow[10]

Méthode et Architecture :

La figure 4 montre le principal flux de données pour cette approche. En effet, afin d'utiliser pleinement les données d'entrée, les flux forward et backward sont calculés pour les images originales I_1 et I_2 , ainsi que leurs images déformées (warped). Une estimation de deux cartes d'occlusion est faite aussi, en vérifiant la cohérence forward-backward. Le modèle de l'enseignant est formé avec une loss photométrique, qui minimise une erreur de déformation en utilisant les 2 images en entrées, les warped images et les carte d'occlusion. Ainsi, ce modèle prédira des estimations de flux optique précises pour les pixels non occlus dans I_1 et I_2 . Pour le modèle d'étudiant, un recadrage est effectué aléatoirement sur I_1 et I_2 . A partir de ses patches, on calcule le flux forward-backward. Une perte photométrique similaire à celle du modèle enseignant est utilisée pour les pixels non occlus dans les patches. En outre, les prédictions du notre modèle d'enseignant sont utilisées en tant qu'annotations de sortie pour guider les pixels occlus dans les patches mais non occlus dans les images originales.

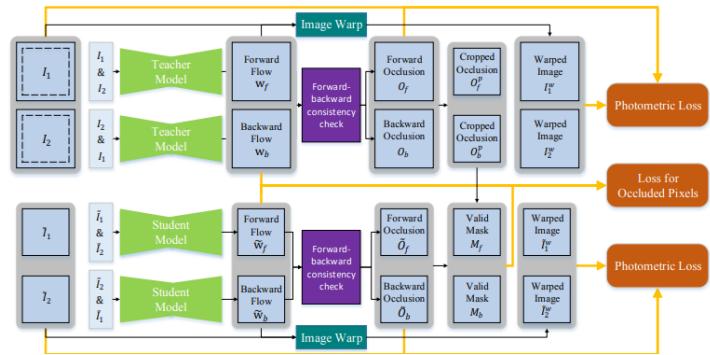


FIGURE 4 – Vue d'ensemble du DDFlow[10]

Le réseau de base de cette approche est le PWC-Net (Sun et al. 2018) en raison de ses performances remarquables et de la taille compacte de son modèle. L'architecture de réseau est identique pour modèle d'enseignant et d'étudiant. La seule différence entre eux est que chacun d'eux a des données d'entrée et des fonctions de perte différentes.

Résultats et critiques :

DDFlow a permis d'atteindre une plus grande précision parmi toutes les méthodes antérieures non supervisées sur tous les datasets de flux optiques difficiles. DDFlow propose bel et bien une approche de distillation des données, qui utilise le recadrage aléatoire pour simuler des occlusions pour l'auto-supervision qui fonctionne particulièrement bien pour les pixels proches des limites de l'image. De ce fait, cette méthode ne peut bien se généraliser pour toutes les occlusions naturelles.

Approche SELF-low [11]

Principe générale de l'approche Self-low

De même comme DDFLOW, SELflow est basée sur une approche de distillation de connaissances. En effet, cette approche entraîne deux CNN (modèle NOC (enseignant) et modèle OCC (étudiant)) avec une même architecture de réseau. Le premier se concentre sur l'estimation précise du flux optique pour les pixels non occlus, et le second apprend à prédire le flux optique pour tous les pixels. Une distillation des estimations de flux non occlus fiables à partir du modèle NOC est réalisé pour superviser l'apprentissage du modèle OCC pour ces pixels occlus. Seul le modèle OCC est nécessaire pour les tests. Au cours de l'apprentissage, des occlusions sont provoquées en perturbant des régions locales avec des bruits aléatoires. Pour se faire, il génère d'abord des superpixels, puis ils sélectionnent plusieurs au hasard à différent endroit et les remplissent de bruit pour couvrir plus de cas d'occlusion. Dans une image cible nouvellement générée, les pixels correspondant aux régions de bruit sont automatiquement occultés. La figure 5 résume bien ce qui a été dit précédemment.

Méthode et Architecture :

Comme on peut le voir dans la figure 6, l'estimation du flux à trois trames dans Self-low est basée sur l'architecture du réseau PWC-Net. Pour mieux comprendre l'approche, le réseau prend trois images en entrée, ce qui produit trois représentations de caractéristiques F_{t-1} , F_t et F_{t+1} . Ensuite, en dehors du flux entrant

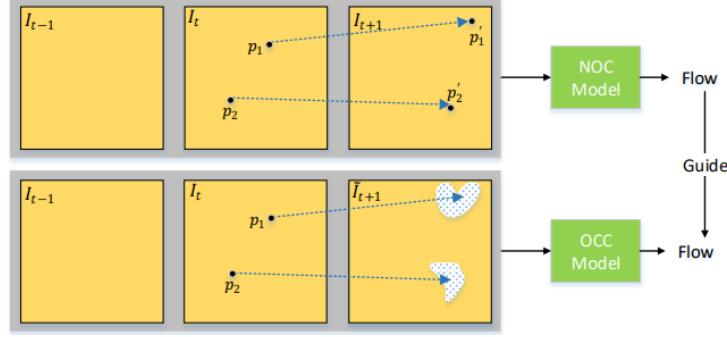


FIGURE 5 – Vue d’ensemble du SELFlow

$w_{t \rightarrow t+1}$ et du volume des coûts entrants, notre modèle calcule également le flux sortant $w_{t \rightarrow t-1}$ et le volume des coûts rétroactifs à chaque niveau simultanément. Pour l'estimation du flux entrant, ils utilisent également les informations sur backward flow et backward cost volume. En effet, la trame antérieur I_{t-1} peut fournir des informations très précieuses surtout pour les régions occultés dans les prochaines frames I_{t+1} mais non occulté dans I_{t-1} . Ce qui fait que l'estimation est plus précise. Le flux forward initial $w_{t \rightarrow t+1}^l$, le négative backward initial flux optique $w_{t+1 \rightarrow t}^l$ caractéristiques de l'image référence F_t^l le volume forward des coûts et le volume backward des coûts sont empilés pour estimer le forward flux à chaque niveau. Pour le backward flux, un échange simple de volume de flux et de coûts comme entrée. Les réseaux d'estimation des flux forward et backward partagent la même structure de réseau et les mêmes poids. Ce réseau simple agrège efficacement les informations temporelles provenant de plusieurs frames pour améliorer la prédiction à bas niveau.

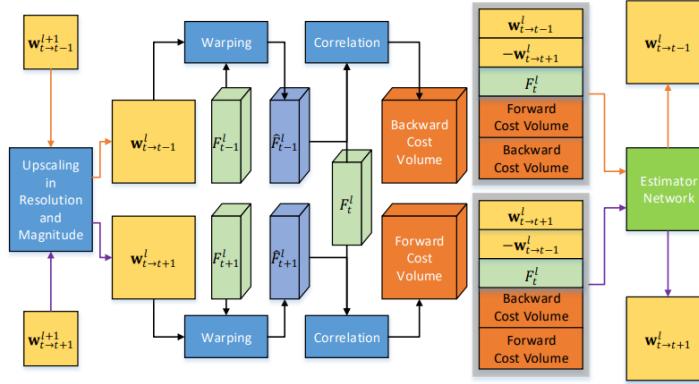


FIGURE 6 – Architecture Globale de SELFLOW

Résultats et critiques

Cette approche a prouvé d'une part que l'utilisation de plusieurs images en entrée peut effectivement améliorer les performances, en particulier pour les pixels occlus du fait qu'ils fournissent plus d'informations, en particulier pour les pixels occlus dans une direction mais non occlus dans la direction inverse. Et d'autre part, que les deux stratégies utilisées pour la simulation d'occlusion : le rectangle et le super-pixel améliorent considérablement les performances, en particulier pour pixels occlus.

Conclusion :

DDFlow et SelFlow utilisent tous deux une approche d'apprentissage en deux étapes pour apprendre les flux optiques de manière autosupervisée. Dans la première étape, ils forment un modèle d'enseignant à l'estimation du flux optique pour les pixels non occlus. Dans un deuxième temps, ils pré-traitent d'abord les images d'entrée, par exemple en recordant et en injectant des bruits de superpixels pour créer des occlusions faites à la main, puis les prédictions du modèle de l'enseignant pour ces pixels non occlus sont considérées comme une vérité de base pour guider un modèle d'étudiant à apprendre le flux optique de pixels faits à la main. Le pipeline général est raisonnable, mais la définition d'occlusion est de manière heuristique. A la première étape, la cohérence forward-backward est utilisée pour détecter si le pixel est occulté. Cependant, cela entraîne des erreurs car de nombreux pixels sont toujours non occultés même s'ils violent ce principe, et vice versa. Il serait plus approprié

de dire que ces pixels sont fiables ou sûrs s'ils passent le contrôle de cohérence forward-backward. De ce point de vue, la création d'occlusions artificielles peut être considérée comme créant des conditions plus difficiles, dans lesquelles la prédiction serait moins sûre. Ensuite, dans la deuxième étape, le point clé est de laisser les prédictions fiables superviser les prédictions moins sûres.

Modèle ARFLOW

C'est une approche d'estimation de flux optique qui modifie le pipeline du processus de l'apprentissage non supervisé classique. On effet, cette méthode effectue une première prédiction non supervisé classique sur les données originales puis effectue une second étape de prédiction en utilisant des données transformés avec la data-augmentation, tout en se servant des prédictions transformées des données originales de la première étape comme signal d'auto-supervision. L'augmentation des données est largement utilisée dans toutes sortes de tâches d'apprentissage. Cependant, la plupart des approches d'estimation de flux optiques non supervisés qu'on a vu dans les sections précédentes n'utilisaient que peu de data augmentation. Du fait, qu'elle soit essentiellement un compromis entre diversité et validité. Elle peut améliorer le modèle en augmentant la diversité des données, tout en entraînant une modification de la distribution des données qui en diminue la précision. De plus l'intégration directe de cette dernière donne de très mauvais résultats en raison de la réduction de la fiabilité de la perte photométrique. Partant de ce point, ARFlow utilise les transformations sur les prédictions des données originales afin de fournir une auto-supervision fiable.

Dans le cadre de ce projet qui vise à implémentée une méthode non supervisé nous avons choisi d'adopter cet approche pour la suite de notre travail.

Méthode :

Le processus du train de cette approche consiste en deux forward (figure 7) : un premier forward sur des échantillons originaux et un second forward sur des échantillons transformés. Le premier forward représente le cadre général des méthodes de flux optique non supervisées, qui Étant donnée deux image I_1, I_2 consécutive, il cherche à prédire le flot optique dense U_{12} qui en l'absence d'une vérité terrain est utilisé pour synthétiser I_1 en modifiant I_2 (warped I_1). Ensuite, une perte photométrique L_{ph} sera calculé en mesurant la différence entre la première image d'entrée et l'image suivante (warped) déformée I_1 en se basant sur l'hypothèse de la constance de la luminosité. Cette L_{ph} n'étant efficace que pour les pixels non occlus. l'idée est alors de créer une carte d'occlusion binaire O_{12} de sorte à désigner ces pixels occlus du coup la perte photométrique dans la région occluse sera éliminée. De plus, pour les endroits sans texture ou avec des motifs répétitifs une régularisation lisse L_{sm} est utilisée afin de contraindre la prédiction à être similaire aux voisins dans les directions x et y lorsqu'il n'y a pas de gradient significatif dans l'image. Pour la seconde forward, des transformations spatiales sur les images, le flux prédit et la carte d'occlusion respectivement pour construire un échantillon augmenté. Comme les prédictions des échantillons transformés sont bruyantes, la carte d'occlusion transformée aux prédictions originales sera intégrée. Ces transformation spatiale fournit une supervision fiable du flux avec un déplacement important ou une occlusion autour de la limite. L'hypothèse de base de cette méthode est que l'augmentation apporte des scènes difficiles dans lesquelles la perte non supervisée ne sera pas fiable, alors que les prévisions transformées des données originales peuvent fournir une auto-supervision fiable. Par conséquent, l'optimisation se fait sur la perte des échantillons transformés L_{aug} plutôt que la perte photométrique L_{ph} du first forward. Après deux retransmissions, la perte photométrique L_{ph} , la régularisation lisse L_{sm} et la régularisation d'augmentation L_{aug} sont rétrogradées en même temps pour mettre à jour le modèle.

*Transformations additionnelles D'autre que la transformation spatiale, une transformation d'occlusion sera réalisé en simulant une occlusion dans le training. Ceci est réalisé en effectuant un recadrage aléatoire qui crée efficacement une nouvelle occlusion dans la limite puis en masquant aléatoirement certains super-pixels dans les images cibles avec du bruit gaussien, ce qui introduit une nouvelle occlusion pour l'image source. La transformation de l'occlusion simplifie la façon de distiller les modèles en optimisant un modèle unique en une seule étape avec un apprentissage de bout en bout.

D'autres transformations ne font que modifier l'apparence des images, comme la gigue aléatoire des couleurs, la luminosité aléatoire, le flou aléatoire, le bruit aléatoire. Cette approche permet d'exploiter ces transformations d'apparence sans aucune contrainte par rapport la loss photométrique puisque cette dernière est calculé dans la first forward.

Architecture

Dans cette approche une architecture légère et étendue à plusieurs cadres basée sur le pipeline principale du *PWC – Net* original est proposé avec quelques modifications afin de réduire encore plus le nombre des paramètres :

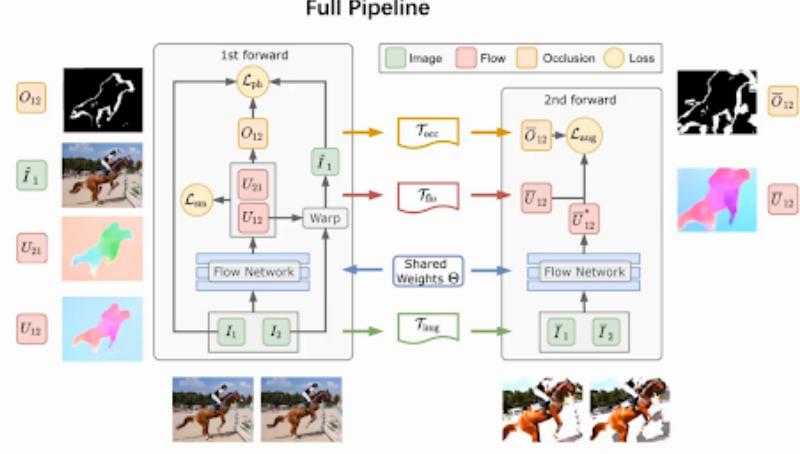


FIGURE 7 – Architecture globale de ARFLOW

- Une connexion complètement dense dans chaque décodeur de PWC-NET est réduite que seules les connexions des deux couches les plus proches sont conservées .
- le décodeur de flux est partagé pour tous les niveaux de la pyramide, avec une couche de convolution supplémentaire pour chaque niveau afin d'aligner les cartes des caractéristiques
- Le modèle est étendu à plusieurs trames en répétant le warping et la corrélation backward des caractéristiques. Le décodeur de flux est partagé à la fois pour le flux forward et le flux backward dans l'extension multi-trames en changeant le signe du flux optique et l'ordre dans la concaténation des caractéristiques.

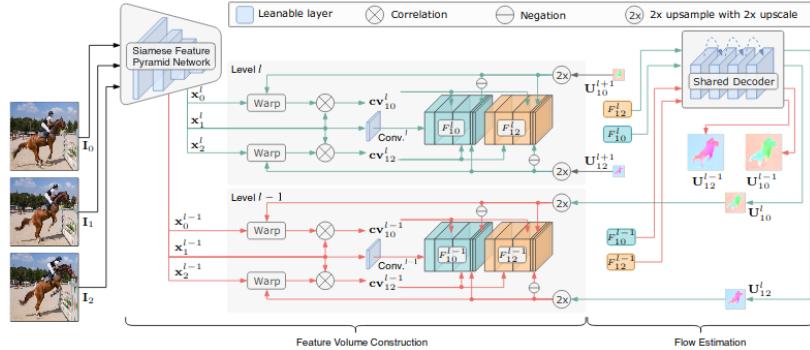


FIGURE 8 – Architecture du modèle multiframe ARWflow

Résultats et conclusion :

Les résultats obtenus à l'issue de l'approche ARFlow montre une amélioration considérablement de la précision, avec une grande compatibilité et une capacité de généralisation. De plus, les observations ont montré que chaque transformation peut améliorer les performances individuellement. Commençant par La transformation spatiale qu'est la plus utile pour toutes les mesures, en particulier pour l'estimation des grands déplacements.La transformation d'occlusion peut améliorer la précision dans la région occultée de manière significative.

Critiques DDFlow, SelFlow et ARFlow :

Nous avons abordé précédemment deux approches basées sur la distillation des connaissances à savoir DDFlow et SELF-flow qui atténuent le problème du manque de fiabilité des objectifs dans les régions occultées. Cependant, nous avons vu que ces méthodes ont été conçues pour le cas d'une occultation partielle seulement. Du coup, le comble serait de pouvoir généraliser la distillation de l'occlusion à d'autres cas de transformation. Pour se faire, dans cette approche diverses transformations sont utilisées pour générer des scènes difficiles telles que la faible luminosité, la surexposition, le grand déplacement ou l'occlusion partielle. Et en guise d'optimisation, un

seul réseau est entraîné dans ARFlow. Ces 3 stratégies entraînent généralement leurs réseau pour fournir des informations préalables "prior" pour faire de l'auto-supervision. Cependant, l'information préalable n'est pas suffisamment précise car un réseau peut facilement être perturbé par des aberrations si la vérité de terrain est inaccessible. En outre, des changements brusques ou des informations préalables imprécises peuvent entraîner une dégradation importante des performances.

Conception et Implémentation

Dans cette partie nous discuterons l'implémentation de ARFlow, nos résultats à l'issue de son implémentations ainsi que les changement apportés au modèle afin d'améliorer les résultats sur notre datasets.

implémentation de ARFlow

Nous avons récupérer l'implémentation officielle sous Pytorch de ARFlow depuis github. Le code a été développé sous Python3, PyTorch 1.1.0 et CUDA 9.0 sur Ubuntu 16.04. Nous avons également récupéré un dossier de modèles pré-entraînés de ARFlow et ARFlow-mv sur KITTI12, KITTI15 et Sintel.

Métriques d'évaluation :

Avant de discuter les résultats nous tenons à préciser les notions liées aux métriques existantes pour estimer la qualités des estimations de flux optique obtenus :

- End-to-end point error (EPE) : elle est définie comme la distance euclidienne entre deux vecteurs, le vecteur de flux optique estimé avec le vecteur de flux optique de vérité de terrain.
- F-score : le F-score ou F-mesure est une mesure de la précision d'un test. Elle est calculée à partir de la précision et du rappel du test, où la précision est le nombre de résultats positifs correctement identifiés divisé par le nombre de tous les résultats positifs, y compris ceux qui ne sont pas correctement identifiés.
- Mesure photo-métrique : Elle a l'avantage de ne pas nécessiter de vérité de terrain. elle est obtenue en utilisant le flux optique pour extrapoler ("warp") la trame courante. L'image warped est ensuite comparée à l'image réelle suivante en utilisant des normes telles que : l1, L2, SSIM ou PSNR.
- norme l1 :La norme L1 qui est calculée comme la somme des valeurs absolues du vecteur.
- Norme l2 : est la norme la plus populaire, également connue sous le nom de norme euclidienne. elle est calculée comme la racine carrée de la somme des carrés des valeurs du vecteur.
- SSIM : Le SSIM mesure la différence perceptive entre deux images similaires
- PSNR : c'est une mesure qui tente de déterminer le niveau de distortion entre deux images .

Dans notre cas à nous et en l'absence de vérité terrain nous avons été contraints d'utiliser la mesure photométrique.

Résultats Générale de l'approche :

Comme nous pouvons le voir dans les deux figures ci-dessous, la comparaison des résultats de ARflow avec ceux des approches supervisées et non-supervisées en utilisant les metrics (AEPE) et pourcentage de pixels erronés (Fl) sur Sintel, KITTI12 et KITTI15 montre que ARFlow surpassé tous les travaux non supervisés précédents avec le moins de paramètres possible. Quant à son extension, cette dernière réduit considérablement l'erreur et permet d'obtenir les meilleures performances avec un minimum de paramètres supplémentaires.

Method	Sintel Training		Sintel Test		# Param.	
	Clean	Final	Clean	Final		
Supervised	FlowNetS-ft [7]	(3.66)	4.44	6.96	7.76	32.07 M
	LiteFlowNet-ft[13]	(1.64)	(2.23)	4.86	6.09	5.37 M
	PWC-Net-ft[38]	(2.02)	(2.08)	4.39	5.04	8.75 M
	IRR-PWC-ft [14]	(1.92)	(2.51)	3.84	4.58	6.36 M
Unsupervised	SelFlow-ft [†] [24]	(1.68)	(1.77)	3.74	4.26	4.79 M
	UnFlow-CSS [27]	-	(7.91)	9.38	10.22	116.58 M
	OccAwareFlow [42]	(4.03)	(5.95)	7.95	9.15	5.12 M
	MFCCoFlow [†] [17]	(3.89)	(5.52)	7.23	8.81	12.21 M
Unsupervised	EpiFlow train-ft [50]	(3.54)	(4.99)	7.00	8.51	8.75 M
	DDFlow [23]	(2.92)	(3.98)	6.18	7.40	4.27 M
	SelFlow [†] [24]	(2.88)	(3.87)	6.56	6.57	4.79 M
	Ours (ARFlow)	(2.79)	(3.73)	4.78	5.89	2.24 M
Ours (ARFlow-MV[†])	(2.73)	(3.69)	4.49	5.67	2.37 M	

FIGURE 9 – I2 : 2006-12-29

Method	KITTI 2012		KITTI 2015		
	training	test	training	test	
Supervised	FlowNet2-ft [15]	(1.28)	1.8	(2.30)	11.48%
	LiteFlowNet-ft [13]	(1.26)	1.7	(2.16)	11.48%
	PWC-Net-ft [38]	(1.45)	1.7	(2.16)	9.60%
	SelFlow-ft[†] [24]	(0.76)	1.5	(1.18)	8.42%
Unsupervised	BridgeDepthFlow [§] [20]	2.56	-	7.02	-
	CCFlow [§] [34]	-	-	5.66	25.27%
	UnOS-stereo [§] [41]	1.64	1.8	5.58	18.00%
	EpiFlow-train-ft [§] [50]	(2.51)	3.4	(5.55)	16.95%
	DDFlow [23]	2.35	3.0	5.72	14.29%
	SelFlow [†] [24]	1.69	2.2	4.84	14.19%
	Ours (ARFlow)	1.44	1.8	2.85	11.80%
Ours (ARFlow-MV[†])		1.26	1.5	3.46	11.79%

FIGURE 10 – Caption

Ces résultats satisfaisants nous ont motivé à tester cette approche sur l'ensemble de nos données.

Datasets :

Comme nous l'avions dit précédemment ARflow a été testé sur 3 différentes datasets à savoir KITTI12[4],KITTI15[12] et Sintel. Les modèles pré-entraînées sont également disponibles.

KITTI : (Karlsruhe Institute of Technology et Toyota Technological Institute) est l'un des ensembles de données les plus populaires pour la robotique mobile et la conduite autonome. ce dataset comporte des images du domaine réel.



FIGURE 11 – Exemple d’Image KITTI

SINTEL : Un ensemble de données pour l’évaluation du flux optique dérivé du court métrage d’animation 3D open source, Sintel.



FIGURE 12 – Exemple d’Image de Sintel

Notre dataset se résume en des images synthétiques de la SST calculées par le logiciel océanographique opérationnel NEMO (Nucleus for European Modeling of the Ocean). Ce logiciel qui fournit une estimation de l’état de l’océan et produit des images SST synthétiques réalistes comme réanalyse obtenue à partir de l’assimilation des données des acquisitions satellitaires avec un modèle physique de l’océan. Ce dataset compte une séquence de 3735 images de (2006-12-28 au 2017-04-06) d’une résolution spatiale de chaque pixel est de 10 kilomètres et l’intervalle de temps entre deux images est d’un jour de taille 384*640. La figure 1 montre certaines de ces images NEMO SST.

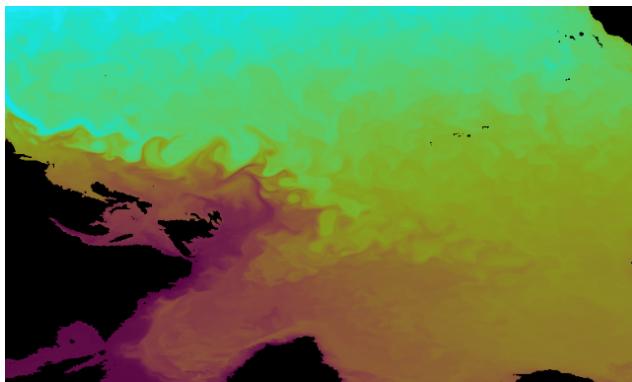


FIGURE 13 – Image du : 2006-12-28

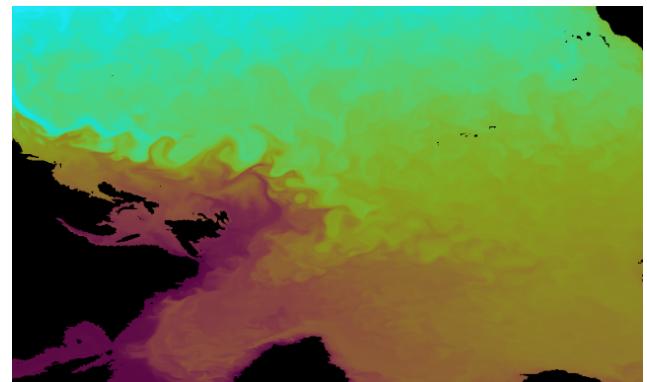


FIGURE 14 – Image du 2007-01-06

Résultats sur notre dataset :

Nous avons d’abord commencé par une inférence des modèles pré-entraînés disponibles avec l’implémentation de ARFLOW sur notre ensemble de données. Nous résumons dans ce qui suit les résultats obtenus à cet égard. Nous tenons à souligner que nous pouvons nullement conclure qu’un résultat est totalement meilleur ou juste en l’absence de vérité terrain, tout ce qui sera conclu est basé sur des métriques qui nous permettent d’estimer à peu près quel résultat est plus satisfaisant. Nous pronons comme exemple deux images consécutives I1 et I2 des dates respectives 2006-12-28 et 2006-12-29 pour afficher les différentes métriques. Les 2 figures ci-dessous représentent cette séquence d’image :

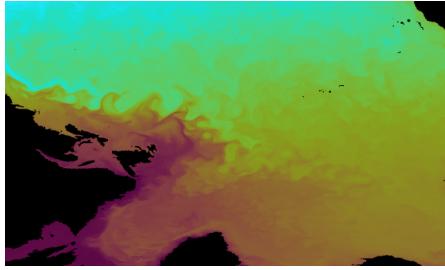


FIGURE 15 – I1 : 2006-12-28

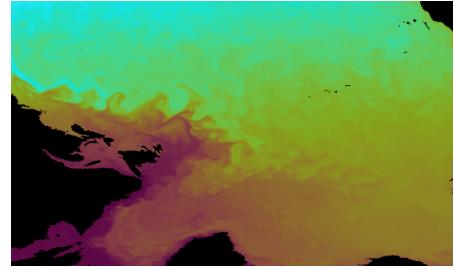


FIGURE 16 – I2 : 2006-12-29

1. Pour le modèle pré-entraîné sur KITTI12

Métriques	résultats
PSNR	44.70595834242519
SSIM	0.9906678312811319
MSE	3.1572577582465278
L1	39.96284315321181

TABLE 1 – Caption

Dans l'ensemble des figures ci-dessus la première figure(28) représente le flux optique obtenu par KITTI 12 pretrained model, la 2ème figure représente l'image I1, la seconde représente la warped I1 avec les 12 plus grande surface de contours de différences entre les deux images.



FIGURE 17 – Flux optique entre I1 et I2 obtenue sous le modèle pré-entraîné sur KITTI12

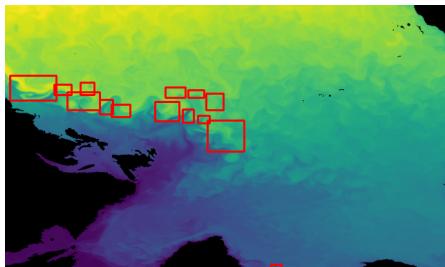


FIGURE 18 – I1 : 2006-12-28

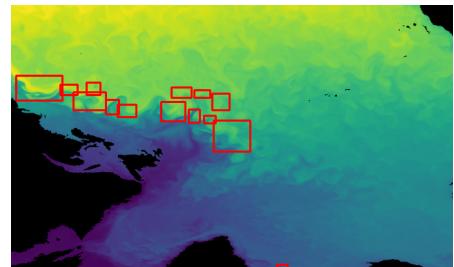


FIGURE 19 – I1 warped

2. Pour le modèle pré-entraîné sur KITTI15 :

Métriques	résultats
PSNR	45.70287425968692
SSIM	0.9922229954346092
MSE	2.3750542534722223
L1	36.52666558159722

TABLE 2 – Caption

3. Pour le modèle pré-entraîné sur Sintel

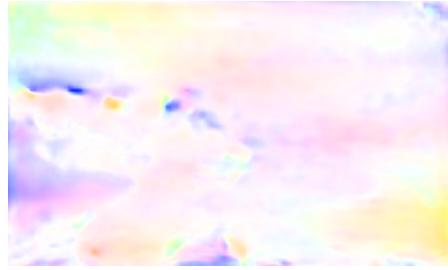


FIGURE 20 – Flux optique entre I1 et I2 obtenue sous le modèle pré-entraîné sur KITTI15

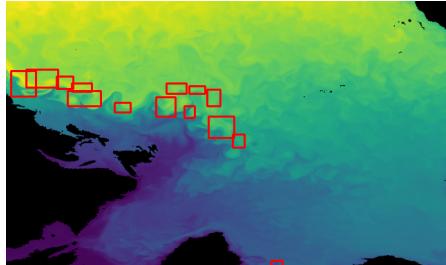


FIGURE 21 – Image I1 : 2006-12-29

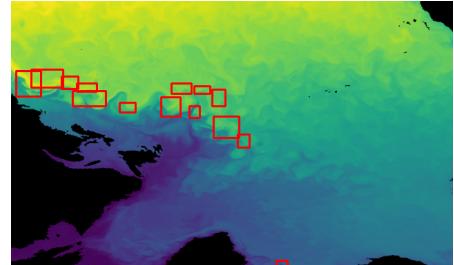


FIGURE 22 – Image I1 warped

Métriques	résultats
PSNR	46.12515816579646
SSIM	0.9930688505398924
MSE	2.212996419270833
L1	35.3597900390625

TABLE 3 – Caption

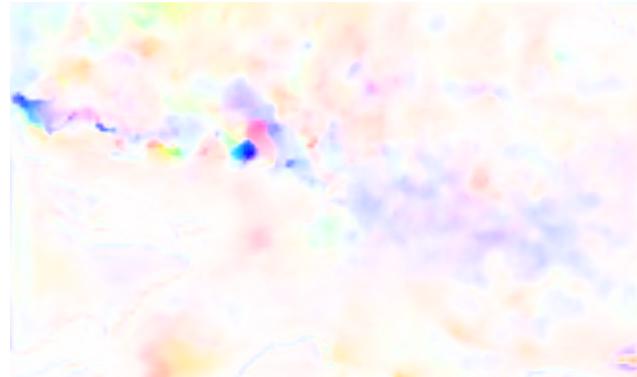


FIGURE 23 – Flux optique entre I1 et I2 obtenue sous le modèle pré-entraîné sur Sintel

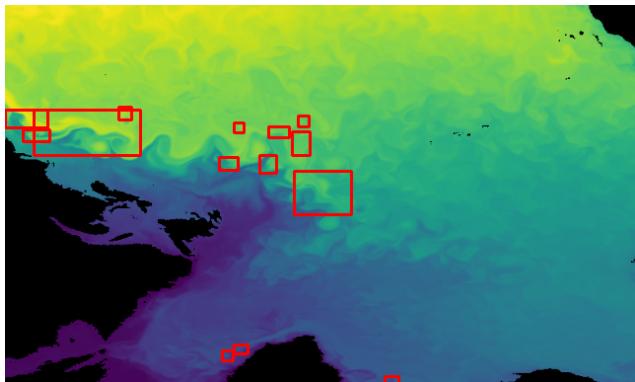


FIGURE 24 – Image I1

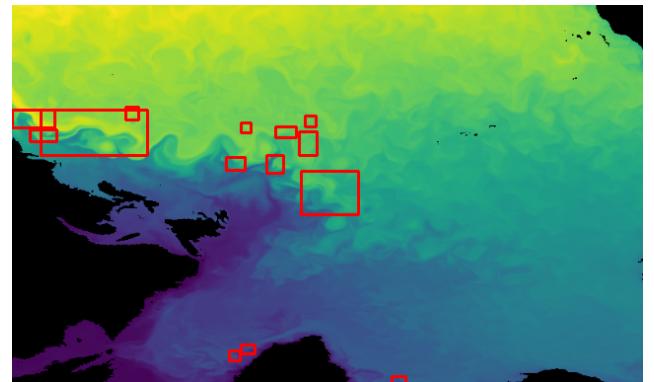


FIGURE 25 – Image I1 warped

Conclusion :

L'ensemble des inférences que nous avons effectué nous ont permis à priori d'avoir des résultats satisfaisants mais beaucoup plus avec Sintel avec de plus haut PSNR et SSIM et de plus bas MSE et L1. On

Métriques	KITTI12	KITTI15	Sintel
PSNR	41.516349623899934	41.65676238999	41.81935533416118
SSIM	0.99434316250688	0.99234366806	0.99349566523666
MSE	6.674010640654842	6.5247895214	6.1420800946316
L1	64.52809387934444	64.3458701886	63.55273053622903

TABLE 4 – Moyenne de chaque métrique sur l’ensemble des données océanographiques

justifie ceci en raison de la similarité du domaine des images Sintel avec ceux de notre datasets contrairement à ceux de kitti qui sont des images du domaine réel. une autre remarque serait de dire que la reconstruction donne des images identique dans les structures comme on peut bel et bien le voir même sur les zones de différence avec un SSIM à 0.99, la différence réside juste dans l’intensité des couleurs. Le tableau ci-dessous résume les résultats obtenus par ces trois modèles sur l’ensemble de nos données.

Contribution :

En se basant sur les résultats qu’on a obtenus précédemment nous avons décidé d’effectuer un transfert learning en utilisant le modèle pré-entraîné sur Sintel afin d’améliorer encore plus nos résultats . Pour se faire nous avons opté à ne garder que la partie encodeur (feature extractor) dans notre nouveau train et on a gelé tout le reste du modèle (décoder). Notre choix se motive du fait que le décoder du modèle sous 1000 epochs a déjà bien appris comment créer un flow depuis des features maps et ça n’a pas vraiment de sens de le modifier car il dépend pas des données. Cependant, les couches du feature extractor nous les avons ré-entraînées sur nos données afin d’extraire les caractéristiques propres à nos données. Nous avons procédé à un train de 50 epochs. en raison de grande limitations par colab, de coupures excessives et du fait que les résultats ne s’améliorent pas autant que ça. Nous allons résumer dans le tableau ci-dessous nos résultats sur l’exemple de séquence d’image précédemment illustré I_1 et I_2 . On effet, dans ces derniers le MSE obtenu est à 52% meilleur que celui de Sintel, et apportent des améliorations même au niveau des autres métriques même si c’est pas aussi considérable que ça.

Métriques	résultats
PSNR	47.855024941487045
SSIM	0.9948284057907755
MSE	1.1731987847222223
L1	33.49497341579861

TABLE 5 – Résultats d’évaluation de OCEANO-prat

Le tableau ci-dessus nous permet de voir l’amélioration obtenue sur l’ensemble des données. L’amélioration n’est pas aussi énorme mais elle reste satisfaisantes.

Métriques	Prat-Oceano
PSNR	42.76912866364636
SSIM	0.9956226148629848
MSE	4.77301263359713
L1	62.756318813169365

TABLE 6 – Moyenne de chaque métrique sur l’ensemble des données océanographiques avec le Modèle transféré Prat-oceano



FIGURE 26 – Flux optique entre I1 et I2 obtenue sous le modèle pré-entraîné sur Notre Modèle

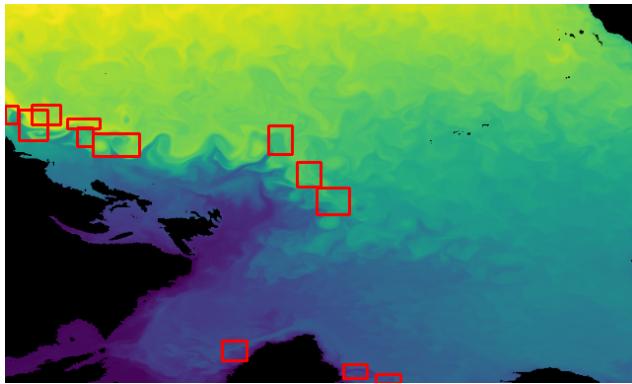


FIGURE 27 – Image I1 : 2006-12-28

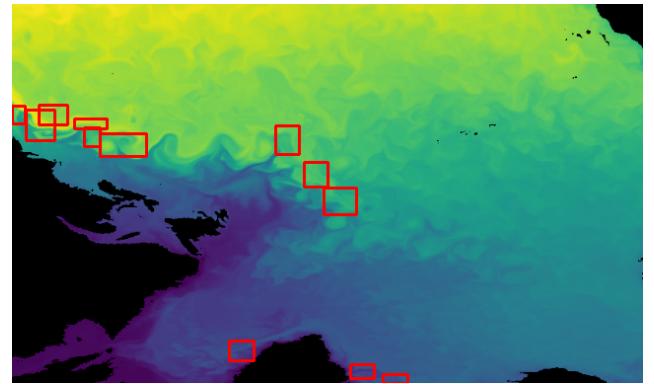


FIGURE 28 – Image I1 warped

OpenCV et Flownet2 :

Nous avons comparé encore nos résultats avec ceux obtenus avec OpenCV et trouvé que encore une fois que nos résultats sont plus bon. Enfin OpenCV a des résultats plus proches à ceux qu'on a obtenu.

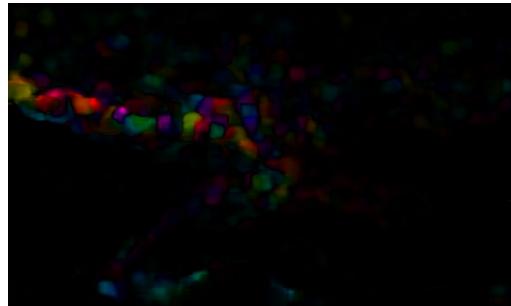


FIGURE 29 – Flux optique entre I1 et I2 obtenue sous OPENCV

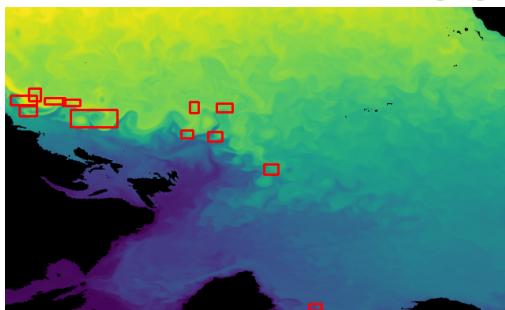


FIGURE 30 – Image I1 : 2006-12-28

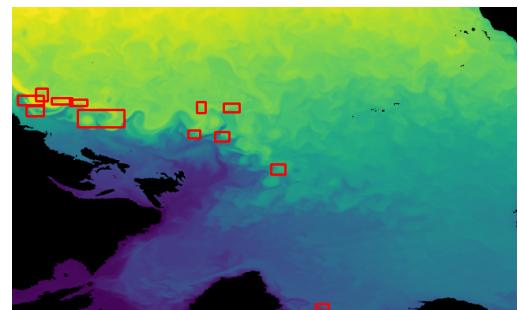


FIGURE 31 – warped Image I1

Pour l'exemple du couple d'image avec OpenCV nous avons obtenu les résultats qu'on résume sur dans le tableau ci-dessous :

Métriques	résultats
PSNR	46.36024916978042
SSIM	0.9938515209127865
MSE	1.855764431423611
L1	35.56558159722222

TABLE 7 – Résultats de l’inférence avec OpenCV

Même chose pour flownet 2, nos résultats dépassent bel et bien les leurs.



FIGURE 32 – Flux optique entre I1 et I2 obtenue sous flownet2

pour l’ensemble de nos données, nous avons obtenu :

Métriques	OpenCV	flownet2
PSNR	41.98129020056284	39.63212866364636
SSIM	0.9930961817669095	0.99106226148629848
MSE	5.706530827727134	6.69018541359713
L1	63.67374021357777	64.6972147169365

TABLE 8 – Moyenne de chaque métrique sur l’ensemble des données océanographiques

On voit bien que notre contribution a apporté une petite amélioration même par rapport aux modèles supervisés.

Conclusion générale

Dans ce modeste travail, nous avons commencé par définir ce que c'est les flots optiques, leur hypothèses et modélisation. Ensuite, nous avons résumait les différentes approches existantes traditionnelles, supervisé et non-supervisé qui traitent du sujet pour avoir une idée globale mais sans pour autant rentrer dans les détails. Nous avons donné par la suite une vue détaillée sur quelques approches de pointes liées à la méthode que nous avons choisi pour la suite de notre travail. Les résultats obtenus par ARFLOW sont en général très satisfaisants et dépassent de loin les méthodes non supervisées et sont aussi performants que des méthodes supervisé. Cette approche se sert de la data-augmentation dans sa prédiction de plus que le processus non supervisé classique.

L'inférence de cette approche sur nos données a donné des résultats satisfaisant en terme de mesure d'évaluation photométriques.

Cependant, afin d'apporter notre touche personnel et d'améliorer les résultats obtenues nous avons effectué un transfert learning sur le modèle. Ceci a été effectuer en ne formant que la partie encodeur, en utilisant Sintel comme modèle pré-entraîné et en gelant la partie décodeur du modèle.

Les résultats se sont effectivement amélioré de point de vue générale, et même que ça donnait des résultats meilleurs sur l'ensemble par rapport à des méthodes supervisées.

Cependant, faute de temps nous n'avons pas exploré la partie multi-frame, on ne peut malheureusement pas en dire autant. Mais nous gardons cette tache comme perspective, et nous vous ferons bien part des résultats qu'on obtiendrons. Nous aimerons aussi explorer d'avantage les possibilités de changement que nous pouvant apporter lors du transfert learning pour améliorer encore plus nos résultats.

Je tiens à signaler que l'approche choisi pour traiter de ce sujet a été modifier après avoir rendu le premier mi-rapport car un autre camarade avait opté pour l'autre approche et d'où les modifications dans l'état de l'art.

Bibliographie

- [1] Aria Ahmadi and Ioannis Patras. Unsupervised convolutional neural networks for motion estimation, 2016.
- [2] S. S. Beauchemin and J. L. Barron. The computation of optical flow. *ACM Comput. Surv.*, 27(3) :433–466, September 1995.
- [3] Philipp Fischer, Alexey Dosovitskiy, Eddy Ilg, Philip Häusser, Caner Hazırbaş, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet : Learning optical flow with convolutional networks, 2015.
- [4] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving ? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012.
- [5] Berthold K.P. Horn and Brian G. Schunck. Determining Optical Flow. In James J. Pearson, editor, *Techniques and Applications of Image Understanding*, volume 0281, pages 319 – 331. International Society for Optics and Photonics, SPIE, 1981.
- [6] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0 : Evolution of optical flow estimation with deep networks, 2016.
- [7] Joel Janai, Fatma Güney, Anurag Ranjan, Michael J. Black, and Andreas Geiger. Unsupervised learning of multi-frame optical flow with occlusions. In *European Conference on Computer Vision (ECCV)*, volume Lecture Notes in Computer Science, vol 11220, pages 713–731. Springer, Cham, September 2018.
- [8] Rico Jonschkowski, Austin Stone, Jonathan T. Barron, Ariel Gordon, Kurt Konolige, and Anelia Angelova. What matters in unsupervised optical flow, 2020.
- [9] Liang Liu, Jiangning Zhang, Ruifei He, Yong Liu, Yabiao Wang, Ying Tai, Donghao Luo, Chengjie Wang, Jilin Li, and Feiyue Huang. Learning by analogy : Reliable supervision from transformations for unsupervised optical flow estimation, 2020.
- [10] Pengpeng Liu, Irwin King, Michael R. Lyu, and Jia Xu. Ddflow : Learning optical flow with unlabeled data distillation, 2019.
- [11] Pengpeng Liu, Michael Lyu, Irwin King, and Jia Xu. Selfflow : Self-supervised learning of optical flow, 2019.
- [12] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3061–3070, 2015.
- [13] Z. Ren, W. Luo, J. Yan, W. Liao, X. Yang, A. Yuille, and H. Zha. Stflow : Self-taught optical flow estimation using pseudo labels. *IEEE Transactions on Image Processing*, 29 :9113–9124, 2020.
- [14] Zhe Ren, Junchi Yan, Bingbing Ni, Bin Liu, Xiaokang Yang, and Hongyuan Zha. Unsupervised deep learning for optical flow estimation. page 1495–1501, 2017.
- [15] Zhe Ren, Junchi Yan, Xiaokang Yang, Alan Yuille, and Hongyuan Zha. Unsupervised learning of optical flow with patch consistency and occlusion estimation. *Pattern Recognition*, 103 :107191, 07 2020.
- [16] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net : Cnns for optical flow using pyramid, warping, and cost volume, 2018.
- [17] L. Tian, Z. Tu, D. Zhang, J. Liu, B. Li, and J. Yuan. Unsupervised learning of optical flow with cnn-based non-local filtering. *IEEE Transactions on Image Processing*, 29 :8429–8442, 2020.
- [18] Bo Yang, Huan Xie, Hongbin Li, Nuohan Li, Anchang Liu, Zhigang Ren, Kuan Ye, Rong Zhu, and Xuezhi Xiang. Unsupervised optical flow estimation based on improved feature pyramid. *Neural Processing Letters*, 08 2020.