Apprentissage de la Cohérence Photométrique pour la Reconstruction de Formes Multi-Vues

Vincent Leroy, Jean-Sébastien Franco, Edmond Boyer

Apprentissage de la Cohérence Photométrique pour la Reconstruction de Formes Multi-Vues

V. Leroy J-S. Franco E. Boyer
INRIA Grenoble Rhônes-Alpes

prenom.nom@inria.fr

Résumé

L'essor des technologies de réalité virtuelle et augmentée s'accompagne d'un besoin accru de contenus appropriés à ces technologies et à leurs méthodes de visualisation. En particulier, la capacité à produire des contenus réels visualisables en 3D devient prépondérante. Nous considérons dans cet article le problème de la reconstruction de scènes 3D dynamiques à partir d'images couleurs. Nous intéressons tout particulièrement à la possibilité de bénéficier des réseaux de neurones convolutifs dans ce processus de reconstruction pour l'améliorer de manière effective. Les méthodes les plus récentes de reconstruction multi-vues estiment des cartes de profondeur par vue et fusionnent ensuite ces cartes dans une forme implicite 3D. Une étape clé de ces méthodes réside dans l'estimation des cartes de profondeurs. Cette étape est traditionnellement effectuée par la recherche de correspondances multi-vues à l'aide de critères de photo-cohérence. Nous proposons ici d'apprendre cette fonction de photo-cohérence sur des exemples au lieu de la définir à travers la corrélation de descripteurs photométriques, comme c'est le cas dans la plupart des méthodes actuelles. L'intuition est que la corrélation de descripteurs d'images est intrinsèquement contrainte et limitée, et que les réseaux profonds ont la capacité d'apprendre des configurations plus larges. Nos résultats sur des données réelles démontrent que cela est le cas. Entraîné sur un jeu de données statiques standard, les réseaux de convolution nous permettent de récupérer des détails sur une forme en mouvement que les descripteurs d'images classiques ne peuvent extraire. Les évaluations comparatives sur ces données standards sont par ailleurs favorables à la méthode que nous proposons.

Mots Clef

Reconstruction Multi Vues, Réseaux de neurones convolutifs, Systèmes Multi-caméras.

Abstract

With the rise of augmented and virtual reality, estimating accurate shapes from multi-view RGB images is becoming an important task in computer vision. The dominant strategy employed for that purpose in the recent years relies on depth maps estimation followed by depth fusion, as depth maps prove to be efficient in recovering local surface details. Motivated by recent success of convolutional neural networks, we take this strategy a step further and present a novel solution for depth map estimation which consists in sweeping a volume along projected rays from a camera, and inferring surface presence probability at a point, seen by an arbitrary number of cameras. A strong motivation behind this work is to study the ability of learning based features to outperform traditional 2D features when estimating depth from multi-view cues. Especially with real life dynamic scenes, containing multiple moving subjects with complex surface details, scenarios where previous image based MVS methods fail to recover accurate details. Our results demonstrate this ability, showing that a CNN, trained on a standard static dataset, can help recovering surface details on dynamic scenes that are not visible to traditional 2D feature based methods. In addition, our evaluation also includes a comparison to existing reconstruction pipelines on the standard evaluation dataset we used to train our network with, showing that our solution performs on par or better than these approaches.

Keywords

Multi-View Stereo Reconstruction, Convolutional Neural Network, Multi-Camera Platform.

1 Introduction

Nous nous intéressons dans cet article au problème de la reconstruction de formes dynamiques réelles à partir de plusieurs vidéos couleurs. Les solutions existantes sont désormais matures et très largement utilisées dans le domaine académique aussi bien qu'industriel. Les applications sont nombreuses et variées que ce soit pour la création de contenus 3D réalistes pour la réalité virtuelle ou augmentée, le divertissement, la préservation de l'héritage culturel, ou encore pour l'étude du mouvement et des déformations dans les applications sportives ou médicales. Un point essentiel et commun à toutes ces applications est la fidélité au réel et en particulier la précision des modèles géométriques construits.

Les méthodes de reconstruction de formes multi-vues at-

teignent aujourd'hui un haut niveau de qualité, et consistent généralement en une extraction de descripteurs, de leurs associations, permettant de discrétiser la forme ou d'extraire un nuage de points, et finalement d'optimiser la géométrie de la surface reconstruite. De manière intéressante, de récents travaux réexaminent le problème d'association stéréo en introduisant des descripteurs et fonctions de similarité inférés à l'aide d'apprentissage profond. De telles solutions permettent d'ajouter un à priori guidé par les données, soit en 2D [42, 24, 43, 41] comme amélioration par rapport aux descripteurs d'images classiques tels que SIFT [23] ou DAISY [37], ou bien en 3D [4, 16, 17], de manière à intégrer la position des caméras, ainsi que des a prioris locaux ou globaux sur les formes observées. Ces nouvelles méthodes on été testées sur des jeux de données multi-vues classiques et leurs performances sont extrêmement prometteuses, obtenant des résultats numériquement meilleurs que les méthodes existantes.

Notre objectif principal est de vérifier que ces améliorations se transférent au cas plus général et complexe d'acquisition de séquences dynamiques. Les challenges apportés par de telles conditions sont généralement une projection des objets d'intérêt plus petite sur les images (dûe au volume de capture plus grand) ; des occultations fréquentes des objets interagissants ; un manque de texture typique des sujets/vêtements réels ; du flou de mouvement dans des actions sport (voir Figure 7). À notre connaissance, aucune technique récente d'apprentissage n'a démontré de capacités de généralisation sur des données dynamiques réelles, mais plutôt en s'évaluant sur les jeux de données d'entraînement (DTU Dataset [15] ou ShapeNet [3]), dans lesquels les problèmes mentionnés ci-dessus ne sont pas correctement représentés.

Parmi toutes les techniques de reconstruction multi vues, des stratégies variées co-existent. Certaines méthodes choisissent de définir des fonctions de disparité symétriques afin d'être agnostique au placement du point de vue [31, 13], tandis que d'autres traitent le problème comme une extraction de carte de profondeurs, où l' observation d'un point de vue de référence doit être corroborée par les autres [25]. Cette dernière permet l'ajout d'un a priori géométrique plus puissant, tout en rendant le système plus robuste aux occultations. En ce qui concerne le champ récepteur du réseau, la majorité des techniques utilise des patches d'image 2D, qui servent à capturer les a priori 2D, mais qui contiennent naturellement moins d'information sur la géométrie 3D inhérente à la scène, et leur utilisation est limitée à un écartement faible entre les caméras [27]. De récents travaux essayent de surpasser ces limitations en raisonnant sur un ensemble de volumes 3D représentant la scène [16] et infèrent l'occupation des voxels d'une grille en parallèle. De telles méthodes sont efficaces pour inférer des formes globales observées dans le jeu de données, mais peuvent aussi rajouter un biais vers des formes particulières.

Nous proposons de traiter le problème de reconstruction comme l'extraction d'une carte de profondeur par caméra,

FIGURE 1 – Scène dynamique complexe, capturée avec un système multi-caméras passif. (*gauche*) une image d'entrée, (*centre*) reconstruction obtenue avec des descripteurs d'image classiques [21], (*droite*) solution proposée. Notre résultat confirme les améliorations que peuvent apporter une approche d'apprentissage, globalement plus robuste dans les zones bruitées ou de faible contraste, telles que les plis de la robe, le bras ou la jambe du modèle.

en gardant néanmoins un support 3D pour l' inférence, afin d'améliorer la précision et la robustesse. Contrairement aux méthodes précédentes, notre volume est défini par la vue référence, de manière à capturer les paramètres de la caméra, et la dépendance 3D inter-vue. Au lieu d' inférer la totalité de l'occupation du volume, nous prédisons un unique score de disparité, pour le centre du volume, facilitant l'entraînement et nous concentrant sur la précision de la détection plutôt que sur une cohérence locale des formes détectées, tout en exploitant la géométrie de la scène. Nous balayons les rayons émergeants d'une caméra à l'aide de notre support volumétrique, nous nommons donc notre méthode *balayage par volume*, et embarquons notre solution dans une méthode typique de reconstruction multi-vues, similaire à [21]. À l'aide de cette stratégie, nous sommes capable de valider l'efficacité de telles méthodes d'apprentissage, obtenant des reconstructions plus précises que les approches classiques. Nos résultats contiennent des détails fins dans des conditions dynamiques complexes, et de meilleure qualité que toutes les solutions antérieures. Nous vérifions aussi que cette méthode bénéficie au cas statique, en l'évaluant sur le DTU Dataset [15], séparé en deux parties, pour l'apprentissage et l'évaluation.

1.1 État de l'art

La reconstruction multi-vues est un problème classique et souvent traité dans la littérature [33, 15]. Alors qu'initialement focalisées sur des scènes statiques, l'application de telles méthodes sur des scènes dynamiques est aujourd'hui un thème de plus en plus populaire. Les approches de reconstruction multi-vues sont une modalité de choix pour les applications de capture de haute fidélité [34, 39, 21, 11, 30, 28], de grands espaces extérieurs

[20, 35, 32], apportant des réponses aux problèmes des méthodes antérieures, basées sur l'enveloppe visuelle, ou capteurs de profondeur [29, 14, 9, 5]. Ces limitations peuvent être : une portée limitée, une sensibilité à l'éclairage, ou une difficulté d'augmenter le nombre de points de vue pour la plupart des technologies, à cause des interférences.

Un grand nombre de représentations ont été introduites pour la capture de séquences, telles que les nuages de points [10], la fusion de cartes de profondeur [25], les maillages [34, 20] ou les discrétisations volumétriques [19, 7, 40]. Néanmoins, toutes ces représentations sont fondées sur un principe commun de photo-cohérence : les rayons émanants d'un point de la surface observée devraient avoir la même apparence, si la surface est visible des caméras. La manière dont est construite cette fonction de disparité fut guidée par les biais qui affectent cette observation, typiquement la géométrie locale de la surface, la distance entre les points de vue, ou la réflectance de l'objet observé.

Dans sa forme la plus simple, cette disparité peut être mesurée en considérant la variance des couleurs observées, telle qu' utilisée dans les travaux précurseurs [19], d'une robustesse limitée. En vision stéréo, avec un écartement inter-vue faible, de simples formes de corrélation 2D normalisées suffisent à caractériser de manière efficace la similarité dans des conditions de contraste et d'éclairage simples. Pour des métriques plus robustes à la géométrie et l' apparence, des descripteurs d'image ont été développés (caractérisations des gradients invariantes à l'échelle) [23, 1, 26], dont certains spécialisés pour l'association dense requise pour la reconstruction multi-vues [37]. Ces solutions ont été appliquées avec succès sur des séquences en mouvement [28, 21]. Dans une certaine mesure, il est possible d'améliorer ces caractérisations en prenant en compte des déformations géométriques locales, par exemple en considérant une distortion homographique guidée par une estimation locale de la normale [10]. La similitude des descripteurs étant caractérisée par une corrélation ou une distance par paires, pour un nombre plus grand de vues, les méthodes de reconstruction considèrent habituellement la mesure de la photo-cohérence soit de manière symétrique dans la combinaison des points de vues [31], ou bien de manière assymétrique, en construisant une carte de profondeur par observation, avec une stratégie de balayage [6, 25]. L'inférence sur des cliques de rayons [40] a été considérée pour embarquer les occultations et co-dépendances le long des rayons. Nous optons pour une stratégie de balayage, qui est généralement plus simple et plus rapide, démontrant néanmoins une robustesse accrue par rapport aux solutions symétriques, notamment aux occultations, fréquentes dans notre cas.

Malgré le succès général des approches de reconstruction, de récents travaux soulignent que de nouveaux a priori et des co-dépendances plus subtiles peuvent encore être appris dans les données réelles. De nombreux travaux étudient la question en apprenant comment comparer des paires de patches 2D [42, 24, 43, 41]. Certains étendent ce principe à la reconstruction multi-vue plus large avec des combinaisons symétriques de descripteurs 2D appris [13]. La limitation commune de ces méthodes en 2D est la difficulté d'ajout, en entrée du réseau, de connaissances sur la scène 3D observée telles que l'orientation locale des rayons, ou la co-tangence de structures de rayons localement planaires. Certains travaux tentent d'apprendre ces a priori 3D, par exemple en estimant simultanément la profondeur et la normale de la surface observée [12].
Nous nous inspirons des méthodes se basant sur un champ récepteur entièrement 3D, permettant de prendre en compte toute forme de correlation 3D [4, 16, 17]. Les principales différences résident dans la forme de notre champ récepteur, construit de manière projective par rapport à la caméra de référence, similairement à certains travaux de stéréo binoculaire [18]. Ceci s'adapte parfaitement à une stratégie de balayage, permettant d'apprendre les correlations 3D relatives à une caméra particulière, sans avoir à apprendre d'une invariance rigide. Cela évite aussi d'avoir à décorréler la résolution de la caméra de la résolution du champ récepteur, l'un étant complètement associé à l'autre. De plus, chaque volume est dédié à l'estimation de l'occupation du point central plutôt que d'inférer la totalité de la grille, simplifiant le problème. Les qualités de cette solution apparaissent dans les reconstructions fines et détaillées de scènes dynamiques complexes, éloignées des jeux de données d'entraînement.

2 Aperçu de la Méthode

Comme beaucoup de solutions existantes, notre méthode consiste à estimer des cartes de profondeur, pour récupérer les détails locaux fins de la surface observée, suivi d'une fusion de cette information permettant d' accumuler les observations, lissant le bruit mais conservant les détails fins cohérents. Nous améliorons cette stratégie avec l'ajout d'une photo-cohérence apprise à l'aide de CNNs. Nous exploitons les capacités des CNN à apprendre des configurations photométriques locales lorsque le point observé est proche de la surface. Comme décrit en Figure 2, notre approche prend en entrée un ensemble d'images calibrées, et reconstruit un maillage 3D correspondant à la surface observée. Les profondeurs le long des rayons passant par les pixels sont obtenues à l'aide d'une stratégie de balayage par volume, consistant en un échantillonnage des profondeurs possibles et la conservation du candidat de probabilité de détection maximale. Pour un point le long d'une ligne de vue, la photo-cohérence est estimée à l'aide d'une discrétisation 3D du volume autour du point. Chaque échantillon du volume contient une paire RGB correspondant à la couleur observée par la caméra de référence, et celle observée par une caméra comparée. Nous collectons ces volumes de paires pour chaque comparaison possible et un CNN est alors utilisé pour détecter les configurations consistantes. Les aspects clés de cette stratégie sont :

— Une approche basée sur la caméra évaluée, qui,

par construction, permet de mieux échantillonner la photo-cohérence en un point qu'une approche globale, pour une meilleure récupération des détails.

— Un support 3D pour l'évaluation de la cohérence, éliminant certaines ambiguïtés amenées par une projection 2D.

— Une stratégie d'apprentissage profond, obtenant de meilleurs résultats que les méthodes précédentes, comme démontré dans nos expériences.

La section suivante décrit la contribution principale du papier : la stratégie d'échantillonnage 3D et l'apprentissage de la photo-cohérence. On peut noter que pour l'étape finale de la méthode, sans perte de généralité, nous utilisons la fonction de distance signée tronquée (FDST) pour fusionner l'information de profondeur des caméras, et nous extrayons la surface de cette forme implicite de la même manière que [21].

3 Estimation de Cartes de Profondeur par Balayage de Volume

Notre approche de reconstruction prend en entrée N images $\{I_i\}_{i=1}^N$, ainsi que leurs opérateurs de projection $\{\pi_i\}_{i=1}^N$, et calcule une carte de profondeur par image, pour finalement fusionner l'information dans une forme implicite 3D. Cette section explique le procédé de construction des cartes de profondeur. Étant donné un pixel p dans une image i, le problème consiste à trouver la profondeur d le long de la ligne de vue, correspondant à l'intersection avec la surface observée. Le point d'intersection le long du rayon partant du pixel p à une profondeur d est noté $r_i(p,d)$. Notre approche consiste à chercher cette profondeur à l'aide d'une fonction de détection, évaluant la probabilité pour un point d'être sur la surface étant donné les couleurs observées par les caméras autour de ce point. Contrairement aux méthodes utilisant des descripteurs ou métriques définies à la main, nous apprenons cette fonction de manière supervisée à l'aide de jeux de données multi-vues, équipés d'une vérité terrain. Dans ce but, nous construisons un réseau de neurones convolutif qui, étant donné une caméra i et un point requête 3D $x \in \mathbb{R}^3$, associe un volume local de paires de couleurs à une décision scalaire, score de photo-cohérence $\rho_i(x) \in [0..1]$. Ce score prend en compte les couleurs vues par la caméra référence à sa résolution native par construction, et rééchantillonne les autres vues dans le volume, pour implicitement prendre en compte l'orientation relative. La nature volontairement assymétrique de cette solution nous permet d'inférer les décisions en prenant en compte les différences de visibilité entre les caméras, c'est à dire de gérer une occultation lorsque les couleurs observées par la caméra de référence ne sont pas confirmées par les autres vues. Ceci étant impossible avec une fonction symétrique telle que [13].
Nous posons donc le problème d'estimation de photo-cohérence comme un problème de classification binaire, à partir des paires de couleurs autour de x, en comparant la caméra i de référence aux autres vues. La section suivante décrit la procédure d' échantillonnage employée, ainsi que les détails de l'architecture du réseau employé. Nous expliquerons finalement le procédé de balayage par volume, appliqué pour extraire les cartes de profondeur.

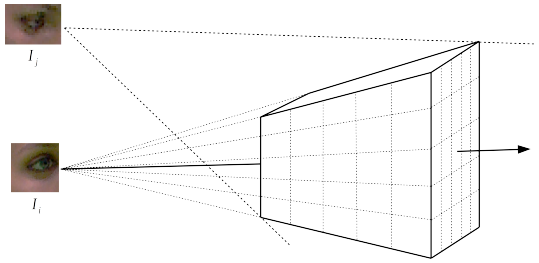

FIGURE 3 – Volume utilisé pour estimer la photo-cohérence le long d'une ligne de vue de l'image de référence i. m^3 échantillons sont régulièrement distribués le long des lignes de vue, et contiennent les paires de couleurs des images i et j. À une profondeur donnée le long du rayon émergeant de i, chaque image $j \neq i$ définit un volume de comparaison par paire.

3.1 Échantillonnage du Volume

Afin d'estimer la photo-cohérence le long d'une ligne de vue, une région d'échantillonnage 3D est déplacée régulièrement le long du rayon. Au sein de cette région, des paires de couleurs reprojetées sur les caméras sont récupérées. Chaque paire contient la couleur de l'image de référence, et celle d'une autre vue. Les échantillons du volume sont espacés régulièrement en profondeur le long des lignes de vue (voir Figure 3). Le volume correspondant est une pyramide tronquée qui se reprojette sur une région de taille définie et constante sur l'image de référence. Cela permet à l'échantillonnage de s'adapter aux propriétés de perception de la caméra, telles que la distance focale, la résolution, la position...
Plus précisément, si on considère la reprojection $r_i(p,d)$, les m^3 échantillons de la grille utilisés pour comparer les images $\{i,j\}_{i \neq j}$ sont alors l'ensemble formé par la projection des pixels d'une fenêtre de taille m^2 de l'image i centrée sur p, échantillonnés régulièrement de la profondeur $d - m\lambda/2$ à la profondeur $d + m\lambda/2$, avec λ choisi tel que l'espacement dans la profondeur soit égal à l'espacement inter-pixel de la caméra de référence, à cette profondeur.
L'échantillonnage est toujours effectué avec la même orientation et ordonnancement par rapport à la caméra de référence. Ainsi, les convolutions sont toujours orientées de manière cohérente par rapport à la direction de profondeur.

Taille du Volume Dans nos expérimentations, et sans perte de généralité, nous choisissons $m = 8$. Notre stratégie est d'apprendre la photo-cohérence le long d'un rayon par paires afin de détecter la présence de la surface, contrairement aux travaux antérieurs [16] essayant d'inférer directement la présence de surface dans des grilles de 32^3

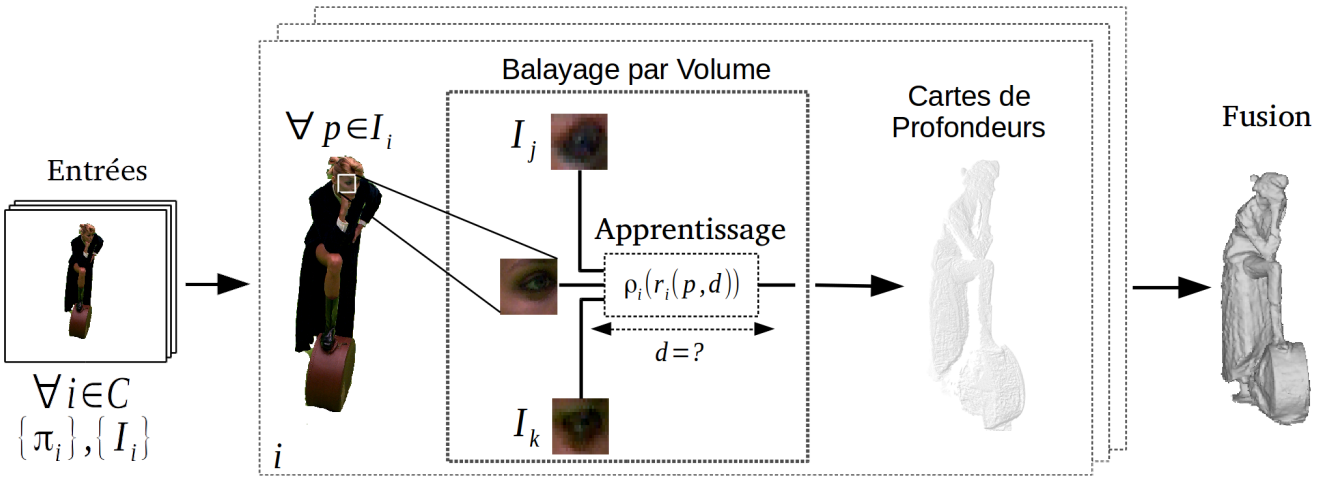

FIGURE 2 – Aperçu de la méthode et notations.

ou 64^3 voxels. En considérant la détection de surface seule comme une tâche bas niveau, et en laissant l'étape de fusion intégrer les profondeurs de manière robuste et consistante, nous simplifions le problème appris ne nécessitant que peu de cohérence spatiale, et autorisant des grilles de plus petite taille.

3.2 Réseau de Neurones Multi-Vues

Comme expliqué précédemment, à un point x le long d'une ligne de vue, on a $N-1$ volumes colorés par des paires de vues, donc $(N-1) \times m^3$ paires de couleurs et nous souhaitons détecter la présence de surface en x. Dans ce but, nous construisons des encodeurs siamois de manière similaire à [13, 42], considérant néanmoins des volumes 3D au lieu de patches 2D d'images. Chaque encodeur construit un descripteur pour un volume de paire. Ce descripteurs sont alors moyennés et nourris à un réseau de décision final. Les réseaux siamois partagent leurs poids, et le moyennage permet au système d'être invariant à l' ordre des caméras. Le réseau est décrit en Figure 4. Les entrées sont $N-1$ volumes colorés de taille $m^3 \times 6$ ou les paires RGB sont concaténées à chaque échantillon. On effectue des convolutions 3D sur les 6 valeurs de cette grille. Les premiers niveaux (encodeur) du réseau prennent les volumes et calculent les descripteurs en parallèle, avec des poids partagés. L'encodeur est une séquence de deux convolutions suivies de non linéarités et de *max-pooling*. Les niveaux de convolution consistent respectivement en 16 et 32 filtres de taille $4 \times 4 \times 4$, suivis d'unités linéaires rectifiées (ReLU) et d'un *max-pooling* de taille $2 \times 2 \times 2$ avec un pas de 2. On calcule ensuite la moyenne des $2 \times 2 \times 2$ descripteurs et calculons la décision finale à l'aide de 128 puis 1 filtres. Le réseau prédit un score $\rho_i(r_i(p,d)) \in [0..1]$ de photo-cohérence à la profondeur d le long du rayon sortant du pixel p de l'image i.

Nous avons expérimenté plusieurs façons d' agréger les descripteurs, et en particulier un *max-pooling*, donnant des résultats de moins bonne qualité. Comparé aux méthodes de [13, 16], le nombre de paramètres du réseau est un à deux ordres de grandeur inférieur. Comme mentionné précédemment, nous pensons que la photo-cohérence est une propriété locale, nécessitant moins de cohérence spatiale que des propriétés de forme.

3.3 Entraînement du Réseau

Le réseau a été entièrement implémenté à l'aide de TensorFlow, et nous avons utilisé les données du DTU Robot Image Dataset [15], jeu de données multi-vues équipé d'une vérité terrain construite à l'aide de lumière structurée, d'une précision d'environ $0.5mm$. De ce jeu de données, nous avons construit 11 million de volumes de comparaisons, avec un nombre de caméras comparées variant aléatoirement de 3 à 40. 80% de ces données ont été utilisées pour l'entraînement, les 20% restant servant à l'évaluation. Nous avons généré un nombre égal d'échantillons positifs et négatifs, en échantillonnant aléatoirement le volume, jusqu'à une distance de $20cm$ de la vérité terrain. Nous avons réalisé l'entraînement en minimisant l'entropie croisée, et en optimisant les poids du réseau par descente de gradient stochastique, en utilisant une estimation du moment adaptative, sur $560,000$ itérations, en optimisant sur 50 grilles simultanément. Les grilles étant relativement petites, et dépendantes du point de vue, nous sommes en mesure de générer assez d'échantillons variés, et n'avons pas besoin de recourir à des techniques d'augmentation.

3.4 Balayage par Volume

Dans le but d'estimer la profondeur le long d'une ligne de vue, notre solution volumétrique s'intègre dans une méthode de balayage par plan standard, en remplaçant le plan par un volume équipé de notre prédicteur. Pour chaque caméra, nous échantillonnons les profondeurs possibles le long du rayon et choisissons le candidat de plus forte probabilité. En pratique, une vue de référence i n'est comparée

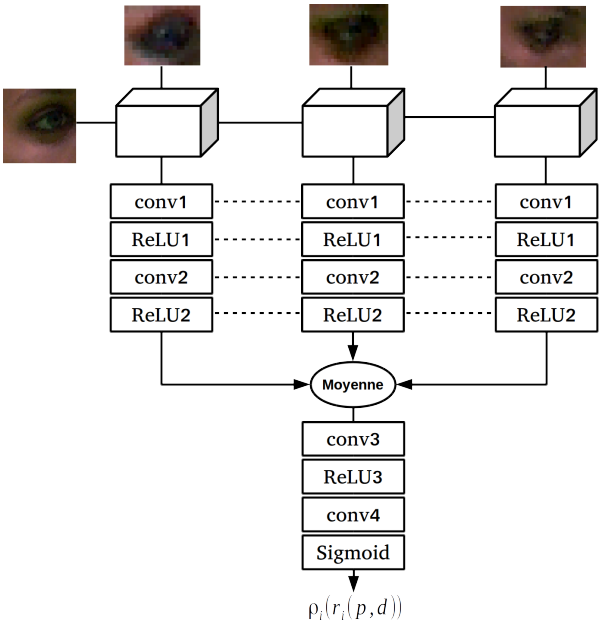

FIGURE 4 – Architecture du CNN. Chaque cube est une comparaison par paire de m^3 échantillons, contenant chacun 6 valeurs, sur lesquels des convolutions 3D sont effectuées. Le score de sortie $\rho_i(r_i(p,d)) \in [0..1]$ encode la photo-cohérence à la profondeur d le long du rayon sortant du pixel p de l'image i.

qu'à ses plus proches voisins, tels que $cos(\phi_{ij}) > 0.7$ où ϕ_{ij} est l'angle entre les axes optiques des caméras i et j. Nous définissons la profondeur estimée d_i^p telle que :

$$d_i^p = \underset{d \in [d_{min}, d_{max}]}{\text{argmax}} \; (\rho_i(r_i(p,d))), \qquad (1)$$

où $\rho_i(r_i(p,d))$ est la mesure de photo-cohérence le long du rayon sortant du pixel p de l'image i. $[d_{min}, d_{max}]$ définit la zone de recherche, qui peut être limitée, en exploitant par exemple l'information d'enveloppe visuelle, si disponible. Les profondeurs sont alors accumulée dans une FDST [8].

4 Résultats

Nous effectuons des évaluations variées afin de quantifier et vérifier les améliorations de notre photo-cohérence multi-vue apprise. Premièrement, nous validons notre approche numériquement dans le cas statique largement étudié, en utilisant le jeu de données de [15]. Nous comparons nos résultats à plusieurs méthodes de reconstruction, basées ou non sur l'apprentissage. Dans un second temps, nous expérimentons notre méthode et sa généralisation dans un cas dynamique drastiquement différent. Nous utilisons pour cela des séquences dynamiques capturées complexes, mettant en évidence les difficultés de telles données, et démontrant une amélioration qualitative comparée aux méthodes récentes [16, 21].

	Précision		Complétude	
Méthode	Moy.	Med.	Moy.	Med.
Tola et al. [38]	0.448	0.205	0.754	0.425
Furukawa et al. [10]	0.678	0.325	0.597	0.375
Campbell et al. [2]	1.286	0.532	0.279	0.155
Ji et al. [16]	0.530	0.260	0.892	0.254
Proposée (*fused*)	1.403	0.422	0.737	0.317
Hartmann et al. [13]	1.563	0.496	1.540	0.710
Proposée (*depthmap*)	1.495	0.430	1.089	0.430

TABLE 1 – Résultats pour toutes les méthodes, sur les 4 objets partagés entre les méthodes [16, 13] (77,110,114,118) de [15] (en mm). Quasiment toutes les méthodes obtiennent des valeurs médianes de l'ordre de la précision de la vérité terrain.

4.1 Evaluation Quantitative

Dans cette section, nous comparons notre solution à de multiples méthodes de reconstruction, sur le DTU Robot Image Dataset [15]. Nous utilisons les métriques de précision et complétude standard afin de quantifier la qualité de la surface estimée. Nous nous comparons aux méthodes [10, 2, 37] fournies avec le jeu de données, ainsi qu'aux méthodes basées sur l'apprentissage [13, 16]. Pour comparer de manière équitable nos résultats à [13], nous comparons uniquement la carte de profondeur calculée sur la même caméra. Afin d'accélérer nos calculs, nous restreignons la recherche de profondeur $10mm$ autour d'une estimation grossière de la profondeur calculée avec les descripteurs DAISY [36]. Le pas de recherche utilisé est $0.5mm$. Comme post-traitement, nous ajoutons simplement un filtre bilatéral doux, prenant en compte la similitude de couleur, la proximité spatiale et la probabilité de détection de la surface. Les résultats des différentes méthodes sont montrés dans le Tableau 1. On obtient des résultats de qualité équivalente aux méthodes existantes en fusionnant 4 points de vue, avec une précision et complétude médiane de l'ordre de la précision de la vérité terrain du jeu de données. De manière intéressante, le réseau utilisé dans [13] possède une architecture similaire, néanmoins basée sur des plans 2D. Ces résultats montrent clairement l'amélioration que peut apporter un support 3D. Comparée à [16], (taille de cube $64 \times 64 \times 64$ et espace inter-échantillon $0.4mm$), on obtient des reconstructions de qualité similaire. Néanmoins, des cas d'erreur non reflétés dans les valeurs moyennes et médianes apparaissent avec leur méthode, qui échoue à gérer certaines occultations et textures répétitives. Par exemple, les plans répétés du toit de la Figure 5.

4.2 Evaluation Qualitative et Généralisation

Afin de vérifier la généralisation de notre méthode à des cas réels complexes, nous effectuons la reconstruction de séquences dynamiques RGB, capturées avec un système multi-caméras complètement différent du jeu d'entraînement, c'est à dire hémisphérique, avec plus de 60 caméras $4K$, dont les distances focales varient de 8 à $25mm$, procurées par les auteurs de [21]. Ce type de système

FIGURE 5 – Comparaison qualitative sur une coupe de l'exemple 29 du DTU [15]. (*gauche*) Nuage de points de la vérité terrain, (*haut droite*) nuage de points de [16], (*bas droite*) Carte de profondeur obtenue avec notre solution.

diffère du cas statique standard comme expliqué en section 1, violant la plupart des a priori de stéréo multi-vue. À cause de ces difficultés, la plupart des méthodes de reconstruction multi-vue échouent sur ce type de données, et des solutions spécifiques doivent être développées, telles que [22, 21]. Nous avons adapté notre algorithme de balayage par volume pour contraindre la recherche de profondeur à l'intérieur de l'enveloppe visuelle, mais avons conservé le réseau entraîné précédemment, sans spécialisation. La Figure 1 montre une reconstruction de notre méthode comparée à [21], conçue pour ce scénario, et basée sur les decripteurs de [36]. Même si leur méthode se comporte correctement dans des scénarios bien contrastés, les descripteurs susmentionnés montrent leurs limites dans des cas peu contrastés et complexes comme celui-ci. La Figure 8 propose la même comparaison dans la région du bras. Un autre exemple est visible en Figure 7, dans laquelle on peut voir que notre solution récupère les détails fins des plis des kimonos, et offre un résultat moins bruité et globalement plus détaillé, particulièrement dans les régions peu contrastées ou occultées.

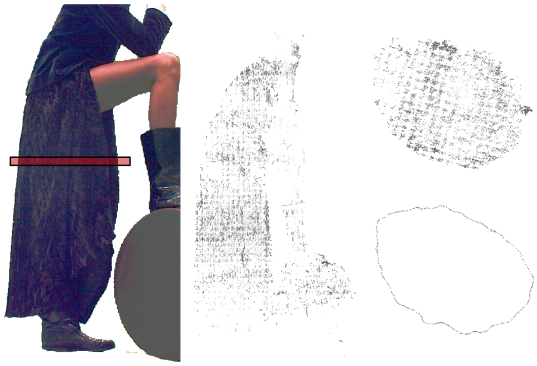

FIGURE 6 – Comparaison qualitative avec [16] dans des condition difficiles. (*gauche*) une image d'entrée, éclaircie, et la tranche visualisée en rouge. (*milieu*) Leur nuage de points reconstruit. (*droite*) Coupe du nuage de points, (*haut*) résultats avec [16], (*bas*) solution proposée.

Nous comparons aussi nos résultats à [16] en Figure 6, en utilisant le code source mis à disposition en ligne par les auteurs. Afin de conduire des expérimentations équitables, nous avons nettoyé les points reconstruits tombant en dehors de l'enveloppe visuelle. Dans ce scénario, les nuages de points obtenus avec leur solution sont extrêmement bruités et difficilement exploitables pour en extraire une surface. On peut observer dans la coupe horizontale des artéfacts provenant de la discrétisation régulière de cette méthode, et sa tendance à reconstruire des nuages de points cohérents. D'un point de vue qualitatif, les résultats obtenus avec notre méthode d'apprentissage démontrent une amélioration drastique, en particulier pour récupérer les détails fins, dans des conditions de bruits extrêmes, et là où les méthodes de reconstructions statiques de détectent presque rien.

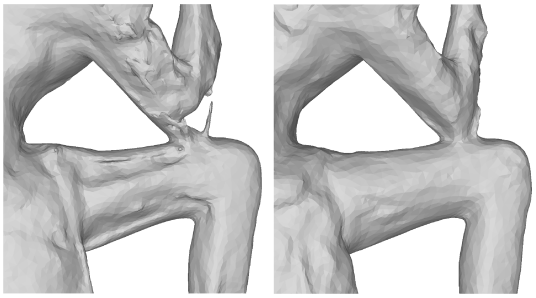

FIGURE 8 – Gros plan sur la région du bras de la Figure 1. (*gauche*) Résultats de [21], (*droite*) notre reconstruction.

5 Conclusion

Nous avons présenté dans ce papier un cadre pour l'apprentissage de la photo-cohérence dans un scénario multi-vues passif. Notre solution consiste à effectuer un balayage de profondeur par volume, dans lequel est reprojeté les couleurs observées par les caméras. Notre système s'adapte à toute configuration de caméras, et grâce à ce nouveau modèle, nous avons validé l'amélioration que les approches d'apprentissage profond peuvent apporter, notamment dans des conditions difficiles, de faible contraste, de bruit et d'occultations. Nous avons obtenu ces résultats avec un nombre de paramètre 10 à 100 fois inférieur aux méthodes précédemment introduites, et montrant des capacités de généralisations impressionnantes, ayant entraîné le modèle sur un jeu de données statiques et appliqué ce même modèle sur des données dynamiques très différentes. Notre méthode permet la récupération de détails plus fins et un niveau de bruit grandement réduit comparé aux méthodes récentes, et offre des perspectives d'applications encore plus complexes.

Références

[1] H. Bay, T. Tuytelaars, and L. J. V. Gool. SURF : speeded up robust features. In *Computer Vision - ECCV 2006, 9th European Conference on Computer Vision, Graz, Austria,*

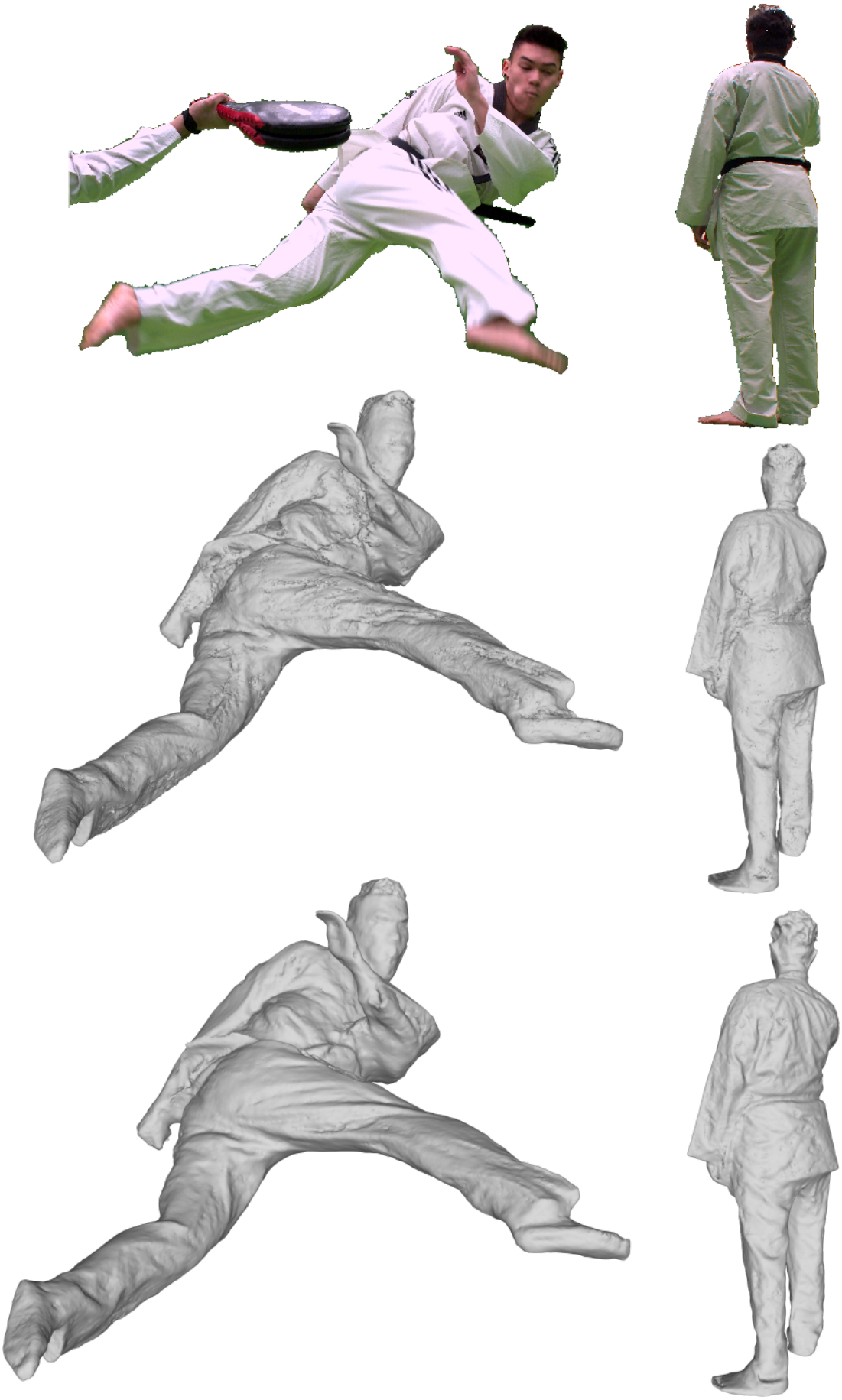

FIGURE 7 – (*haut*) Scène dynamique réelle complexe, contenant des mouvements rapides et un contraste faible. (*milieu*) Reconstruction de [21], (*bas*) résultats obtenus avec notre méthode.

May 7-13, 2006, Proceedings, Part I, pages 404–417, 2006. 3

[2] N. D. F. Campbell, G. Vogiatzis, C. Hernández, and R. Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *Computer Vision - ECCV 2008, 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I*, pages 766–779, 2008. 6

[3] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. ShapeNet : An Information-Rich 3D Model Repository. Technical Report arXiv :1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 2

[4] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3d-r2n2 : A unified approach for single and multi-view 3d object reconstruction. In *ECCV*, 2016. 2, 3

[5] A. Collet, M. Chuang, P. Sweeney, D. Gillett, D. Evseev, D. Calabrese, H. Hoppe, A. G. Kirk, and S. Sullivan. High-quality streamable free-viewpoint video. *ACM Trans. Graph.*, 34(4) :69, 2015. 3

[6] R. T. Collins. A space-sweep approach to true multi-image matching. In *CVPR*, pages 358–363. IEEE Computer Society, 1996. 3

[7] D. Cremers and K. Kolev. Multiview stereo and silhouette consistency via convex functionals over convex domains. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(6) :1161–1174, June 2011. 3

[8] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1996, New Orleans, LA, USA, August 4-9, 1996*, pages 303–312, 1996. 6

[9] M. Dou, S. Khamis, Y. Degtyarev, P. Davidson, S. R. Fanello, A. Kowdle, S. O. Escolano, C. Rhemann, D. Kim, J. Taylor, P. Kohli, V. Tankovich, and S. Izadi. Fusion4d : Real-time performance capture of challenging scenes. *ACM Trans. Graph.*, 35(4) :114 :1–114 :13, July 2016. 3

[10] Y. Furukawa and J. Ponce. Accurate, dense, and robust multi-view stereopsis. In *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA*, 2007. 3, 6

[11] J. Gall, C. Stoll, E. D. Aguiar, C. Theobalt, B. Rosenhahn, and H. peter Seidel. Motion capture using joint skeleton tracking and surface estimation. In *CVPR*, 2009. 2

[12] S. Galliani and K. Schindler. Just look at the image : Viewpoint-specific surface normal prediction for improved multi-view reconstruction. In *CVPR*, pages 5479–5487. IEEE Computer Society, 2016. 3

[13] W. Hartmann, S. Galliani, M. Havlena, L. Van Gool, and K. Schindler. Learned multi-patch similarity. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2, 3, 4, 5, 6

[14] M. Innmann, M. Zollhöfer, M. Nießner, C. Theobalt, and M. Stamminger. Volumedeform : Real-time volumetric non-rigid reconstruction. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, pages 362–379, 2016. 3

[15] R. R. Jensen, A. L. Dahl, G. Vogiatzis, E. Tola, and H. Aanæs. Large scale multi-view stereopsis evaluation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 406–413, 2014. 2, 5, 6, 7

[16] M. Ji, J. Gall, H. Zheng, Y. Liu, and L. Fang. Surfacenet : An end-to-end 3d neural network for multiview stereopsis. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 2, 3, 4, 5, 6, 7

[17] A. Kar, C. Häne, and J. Malik. Learning a multi-view stereo machine. In *Proc. Neural Information Processing Systems (NIPS)*, 2017. 2, 3

[18] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017. 3

[19] K. N. Kutulakos and S. M. Seitz. A theory of shape by space carving. *International Journal of Computer Vision*, 38(3) :199–218, Jul 2000. 3

[20] P. Labatut, J. Pons, and R. Keriven. Efficient multi-view reconstruction of large-scale scenes using interest points, delaunay triangulation and graph cuts. In *IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007*, pages 1–8, 2007. 3

[21] V. Leroy, J.-S. Franco, and E. Boyer. Multi-View Dynamic Shape Refinement Using Local Temporal Integration. In *IEEE, International Conference on Computer Vision 2017*, Venice, Italy, Oct. 2017. 2, 3, 4, 6, 7, 8

[22] Y. Liu, Q. Dai, and W. Xu. A point-cloud-based multiview stereo algorithm for free-viewpoint video. *IEEE Trans. Vis. Comput. Graph.*, 16(3) :407–418, 2010. 7

[23] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2) :91–110, 2004. 2, 3

[24] W. Luo, A. G. Schwing, and R. Urtasun. Efficient deep learning for stereo matching. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5695–5703, 2016. 2, 3

[25] P. Merrell, A. Akbarzadeh, L. Wang, J. michael Frahm, and R. Y. D. Nistér. Real-time visibility-based fusion of depth maps. In *Int. Conf. on Computer Vision and Pattern Recognition*, 2007. 2, 3

[26] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2003), 16-22 June 2003, Madison, WI, USA*, pages 257–263, 2003. 3

[27] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(10) :1615–1630, Oct. 2005. 2

[28] A. Mustafa, H. Kim, J. Guillemaut, and A. Hilton. Temporally coherent 4d reconstruction of complex dynamic scenes. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4660–4669, 2016. 2, 3

[29] R. A. Newcombe, D. Fox, and S. M. Seitz. Dynamicfusion : Reconstruction and tracking of non-rigid scenes in real-time. In *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR) 2015, Boston, MA, USA, June 7-12, 2015*, pages 343–352, 2015. 3

[30] M. R. Oswald and D. Cremers. A convex relaxation approach to space time multi-view 3d reconstruction. In *ICCV Workshop on Dynamic Shape Capture and Analysis (4DMOD)*, 2013. 2

[31] J.-P. Pons, R. Keriven, and O. Faugeras. Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score. *IJCV*, 72(2) :179–193, 2007. 2, 3

[32] T. Schöps, J. L. Schönberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 3

[33] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, pages 519–528, 2006. 2

[34] J. Starck and A. Hilton. Surface capture for performance-based animation. *IEEE Comput. Graph. Appl.*, 27(3) :21–31, May 2007. 2, 3

[35] C. Strecha, W. von Hansen, L. V. Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. 3

[36] E. Tola, V. Lepetit, and P. Fua. A fast local descriptor for dense matching. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA*, 2008. 6, 7

[37] E. Tola, V. Lepetit, and P. Fua. DAISY : an efficient dense descriptor applied to wide-baseline stereo. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(5) :815–830, 2010. 2, 3, 6

[38] E. Tola, C. Strecha, and P. Fua. Efficient large-scale multi-view stereo for ultra high-resolution image sets. *Mach. Vis. Appl.*, 23(5) :903–920, 2012. 6

[39] T. Tung, S. Nobuhara, and T. Matsuyama. Complete multi-view reconstruction of dynamic scenes from probabilistic fusion of narrow and wide baseline stereo. In *IEEE 12th International Conference on Computer Vision, ICCV 2009, Kyoto, Japan, September 27 - October 4, 2009*, pages 1709–1716, 2009. 2

[40] A. O. Ulusoy, A. Geiger, and M. J. Black. Towards probabilistic volumetric reconstruction using ray potentials. In *3D Vision (3DV), 2015 3rd International Conference on*, pages 10–18, Lyon, Oct. 2015. 3

[41] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox. Demon : Depth and motion network for learning monocular stereo. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5622–5631, 2017. 2, 3

[42] J. Žbontar and Y. LeCun. Stereo matching by training a convolutional neural network to compare image patches. *J. Mach. Learn. Res.*, 17(1) :2287–2318, Jan. 2016. 2, 3, 5

[43] S. Zagoruyko and N. Komodakis. Learning to compare image patches via convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4353–4361, 2015. 2, 3