Quelle image met le mieux en valeur un modèle 3D ?

Marie Pelissier-Combescure, Géraldine Morin, Sylvie Chambon

Quelle image met le mieux en valeur un modèle 3D ?

M. Pelissier-Combescure[1] G. Morin[2] S. Chambon[3]

[1] Université de Toulouse, IRIT – Toulouse INP, France
{marie.pelissier-combescure, geraldine.morin, sylvie.chambon}@irit.fr

FIGURE 1 – Classement automatique obtenu par l'approche proposée pour un ensemble d'images réelles en fonction de leur pertinence vis-à-vis d'un objet d'intérêt (dans cet exemple le canapé, du plus pertinent au moins pertinent, de la gauche vers la droite).

Résumé

Étant donné une image d'un objet d'intérêt, nous souhaitons quantifier la qualité de la vue de cet objet 3D dans cette représentation 2D. Pour cela, nous définissons un score de pertinence permettant d'ordonner un ensemble d'images : de celle mettant le mieux en valeur l'objet à celle présentant la plus mauvaise mise en valeur. Ce score s'appuie sur trois critères complémentaires relatifs à l'objet étudié dans l'image : la dominance, la taille et la quantité d'information caractéristique disponible sur cet objet, dans l'image. Cette information caractéristique exploite les résultats fournis par un détecteur de saillance curviligne. De manière complémentaire, nous considérons également les scores de confiance fournis par la sortie d'un réseau de neurones (nous choisissons des réseaux neurones de référence dans le domaine de la détection ou de la classification). Afin de valider et de comparer l'approche introduite et les méthodes étudiées, nous avons mis en place un protocole utilisant des images permettant de proposer un classement de référence. Nos résultats expérimentaux démontrent l'efficacité de notre méthode et permettent de comprendre le comportement des réseaux de neurones. Nous fournissons également des résultats qualitatifs visuels sur des jeux de données réels pour illustrer l'intérêt de l'approche.

Mots Clef

2D/3D, saillance, points d'intérêt, apprentissage profond, *score de pertinence*.

Abstract

Given an object present in an image, our purpose is to quantify how well this object is represented in this view. To do this, we define a highlighting score that allows us to rank a set of images : from the one that best showcases the object to the worst one. To quantify the showcase of an object in the image, we combine three complementary criteria into a highlighting score : its dominance, its size and the quantity of its characteristic information based on a score given by a curvilinear saliency detector. As an alternative, we consider the confidence scores of state of the art detection and classification neural networks. In order to validate the proposed approaches based on these scores, we provide a validation protocol based on a set images we generate to provide a reference classification. Our experimental results demonstrate the efficiency of our method and help understanding the behaviour of the networks. We also illustrate the interest of the approach with visual qualitative results on a real dataset.

Keywords

2D/3D, saliency, interest points, deep learning, highligthing score.

1 Introduction

Il est de plus en plus simple d'obtenir des données visuelles massives car elles sont de plus en plus facile à générer. En conséquence, des bases de données proposent des milliers d'images, comme Imagenet [1], ou des dizaines d'heures de vidéos, comme ToCaDa [2]. Le jeu de don-

FIGURE 2 – Différence entre un point de vue favorable (colonne de gauche) et un point de vue défavorable (colonne de droite). En (b), le canapé est vu de dos, donc tous les détails caractéristiques ne sont pas visibles contrairement à l'orientation du canapé en (a). Dans (d), le canapé est caché par de nombreuses occultations et, dans (f), le canapé n'est pas l'objet central de l'image.

nées Pascal3D+ [3] contient un ensemble d'images, avec de multiples annotations, ainsi que des modèles 3D des objets présents dans les images. Ces objets peuvent apparaître au premier plan de l'image ou faire partie de l'arrière-plan de l'image. En outre, les objets peuvent être tronqués, occultés ou ne pas être mis en évidence dans l'image. Néanmoins, ces grands volumes de données (images, vidéos, modèle 3D ou cartes de profondeur) contiennent des informations plus ou moins pertinentes et importantes.

L'objectif de cet article est de quantifier la qualité de la représentation 2D d'un objet présent dans une image, comme l'illustre la figure 2 où le canapé est toujours mieux représenté dans la première colonne que dans la seconde. Plus précisément, étant donné un ensemble d'images d'un objet d'intérêt, la méthode proposée est capable de classer les images en fonction de la qualité de la mise en valeur de l'objet, c'est-à-dire de la meilleure à la pire représentation suivant un score de mise en valeur nommé *score de pertinence*, comme illustré sur la figure 1. Déterminer l'image la plus pertinente pour représenter un objet est utile à de nombreuses applications, allant de la médecine aux voitures autonomes. Par exemple, dans les applications médicales de réalité augmentée, les images de la caméra sont enregistrées en temps réel par rapport au modèle 3D d'un organe donné dans un paradigme de suivi par détection [4] et il serait intéressant de présenter la vue la plus pertinente pour le chirurgien. Dans un contexte de conduite autonome, les objets d'intérêt, comme un piéton, pourraient être identifiés à partir de l'image qui les met le mieux en valeur [5].

La mise en valeur de l'objet dépend de son placement dans l'image. En effet, la position de l'objet peut être plus ou moins favorable ou significative, puisqu'elle influence la quantité d'informations caractéristiques de l'objet que nous pouvons extraire. De plus, la mise en évidence de l'objet dépend de la complexité et de la densité de la scène dans laquelle l'objet est projeté dans l'image. Afin de mesurer le *score de pertinence*, nous avons besoin d'une référence pour cet objet et nous choisissons de le caractériser par sa géométrie, c'est-à-dire un maillage 3D sans texture. Nous supposons les paramètres intrinsèques et extrinsèques de la caméra connus et ainsi nous sommes capables de placer les silhouettes 3D dans la même position que celles qu'elles occupent dans les images 2D.

Comme nous souhaitons bénéficier des informations fournies par la texture dans l'image 2D et par la géométrie dans le maillage 3D, nous proposons dans un premier temps une méthode déterministe s'appuyant sur un outil classique de la vision : l'extraction des primitives d'intérêt. De plus, au cours des dix dernières années, les réseaux de neurones sont de plus en plus utilisés et ont montré leur efficacité dans de nombreux domaines, comme la classification ou la régression. C'est pourquoi, dans un second temps, nous utilisons également les scores de confiance produits par les réseaux de neurones pour quantifier la mise en valeur d'un objet dans une image.

Après avoir brièvement présenté les méthodes existantes pour évaluer la qualité des images, nous passons en revue les approches pour la détection des primitives saillantes en 2D et en 3D, section 2, ainsi que les approches récentes s'appuyant sur l'utilisation de réseaux neuronaux, section 3. La section 4 présente l'algorithme proposé avant de détailler les différents scores proposés et étudiés. Plus précisément, la section 5 décrit le *score de pertinence* déterministe tandis que la section 6 présente le score de confiance obtenu en sortie des différents réseaux de neurones. Les performances de ces scores sont mesurées par une étape de validation objective décrite dans la section 7. L'analyse de ces performances est donnée dans la section 8. Enfin, la section 9 montre l'efficacité de notre approche à classer avec succès des images réelles en fonction de leur capacité à mettre en valeur un objet.

2 Qualité d'une vue 2D à partir de primitives saillantes

L'évaluation de la qualité d'une image dans son ensemble est un concept très subjectif. En effet, il existe de

nombreuses mesures de similarité ou d'évaluation de la qualité mais il est difficile de déterminer le lien avec la perception visuelle humaine. Une approche classique consiste à évaluer la qualité de l'image en mesurant le niveau de flou ou la présence de bruit, comme dans [6]. Certains travaux utilisent les informations colorimétriques, comme dans [7]. Plus récemment, en faisant intervenir des réseaux de neurones, les auteurs de [8] déterminent la combinaison la plus harmonieuse de meubles 3D pour la décoration d'une pièce du point de vue du style. Ils mesurent à la fois la compatibilité et les similitudes entre différents styles de meubles. Le travail que nous proposons est complémentaire au leur puisque nous souhaitons qualifier l'intérêt d'une vue relative à un objet 3D en évaluant comment la disposition ou la position de l'objet et le contexte de la scène mettent en valeur cet objet particulier. Pour cela, nous nous appuyons sur une information différente : la présence de points caractéristiques.

En 2D, de nombreux détecteurs ont été proposés, et le plus classique, le détecteur de Harris [9] est du premier ordre, c'est-à-dire qu'il s'appuie sur les dérivées premières des images. Une autre catégorie de détecteurs utilise la notion de région [10, 11, 12], c'est-à-dire qu'il existe une région caractéristique autour du point étudié qui permet de déterminer son unicité. Une dernière famille de détecteurs fait intervenir les dérivées secondes et, plus précisément, la courbure [13]. Plus récemment, la notion d'analyse multi-échelle a été introduite pour chacun de ces types de détecteur et l'approche la plus connue est SIFT, *Scale Invariant Features Transform* [14]. En 3D, nous trouvons également des primitives de premier ordre, de second ordre et utilisant la courbure. La plupart des techniques reposent sur la généralisation des détecteurs de la 2D à la 3D, comme le montrent les nombreuses publications qui visent à généraliser le détecteur SIFT à la 3D, comme par exemple les travaux de [15].

Notre objectif nécessite un détecteur capable de mettre en avant des caractéristiques saillantes répétables, à la fois en 2D et en 3D. La principale difficulté réside dans la nature des données manipulées. En effet, une image 2D présente une texture, un fond et un éclairage que l'on ne retrouve pas dans le modèle 3D non texturé. Le détecteur [16] s'appuie sur la notion de courbure tout en diminuant l'effet de la texture (grâce à une analyse multi-échelle et la prise en compte du concept de carte de focalisation). Les résultats présentés dans [16] montrent que ce détecteur permet la meilleure répétabilité des points extraits sur les différentes modalités par rapport aux détecteurs de l'état de l'art tels que SIFT. Ainsi, dans cet article, nous proposons d'utiliser ce détecteur pour caractériser et comparer l'objet dans les deux modalités.

3 Réseaux de neurones

Depuis 2010, de nombreuses avancées (meilleure formalisation, augmentation des bases de données annotées et de la puissance de calculs) ont permis le succès de l'apprentissage profond tel que nous le connaissons aujourd'hui. En ce qui concerne le traitement d'images, c'est l'introduction du défi *ImageNet*, débuté en 2010, et l'arrivée du réseau *AlexNet* [17] en 2012 qui ont initié l'utilisation massive de l'apprentissage profond dans ce domaine sous la forme de réseaux de neurones convolutifs *Convolutional Neural Network*, CNN.

Les réseaux les plus proches de la question que nous nous posons dans ce papier sont ceux entraînés pour évaluer l'esthétique d'une image [18, 19]. La prédiction de la qualité esthétique peut s'appuyer à la fois sur des informations globales et locales comme dans [20]. Une autre possibilité est d'introduire un mécanisme d'attention [21]. Tous ces travaux permettent de fournir une qualité globale d'une image alors que nous nous intéressons à la qualité relative à la mise en valeur d'un objet particulier de la scène.

En ce qui concerne les problèmes de classification, c'est-à-dire les réseaux de neurones qui estiment une classe pour chaque image d'entrée, les réseaux les plus connus sont : la version simplifiée de *AlexNet* : *VGG-16* [22], le réseau *Inception* [23], le réseau *Resnet* [24] qui a remporté le challenge *Imagenet* et le plus récent *EfficientNet* [25].

Depuis 2014, des architectures spécialement adaptées à la détection d'objets ont également été introduites, comme : *Region based CNN*, *R-CNN* [26], et ses variantes [27, 28], *Single Shot Detector*, *SSDs* [29], et *You Only Look Once*, *YOLO* [30]. Pour une image d'entrée donnée, les sorties de ces réseaux sont les coordonnées de la boîte englobante et les probabilités de l'étiquette de classe correspondantes à chaque objet détecté dans l'image. Ils traitent la détection d'objets comme un problème de régression. Les détecteurs de la première famille, *R-CNN*, sont peu efficaces et, en conséquence, les réseaux *SSDs* et *YOLO* ont été créés afin de diminuer la complexité.

4 Méthode proposée

Nous souhaitons quantifier la qualité d'une représentation 2D d'un objet 3D, c'est-à-dire, à partir d'un ensemble d'images contenant le même objet 3D, identifier les images les plus révélatrices, cf. figure 3. En d'autres termes, les images qui mettent le mieux en valeur l'objet, c'est-à-dire qui mettent en évidence ses éléments caractéristiques. Ainsi, pour déterminer la qualité du point de vue, nous proposons d'utiliser les informations saillantes extraites dans chaque type de données. De plus, intuitivement, l'image sera révélatrice si l'objet n'est pas occulté ou tronqué et s'il apparaît au premier plan. Enfin, il est préférable qu'il soit placé dans une scène peu complexe afin de réduire le risque que d'autres objets de la scène ne soient plus proéminents que lui.

Pour comparer ces données, il est d'abord nécessaire de déterminer une représentation commune. Dans la littérature, un choix classique consiste à projeter le modèle 3D depuis différents points de vue (généralement sur une

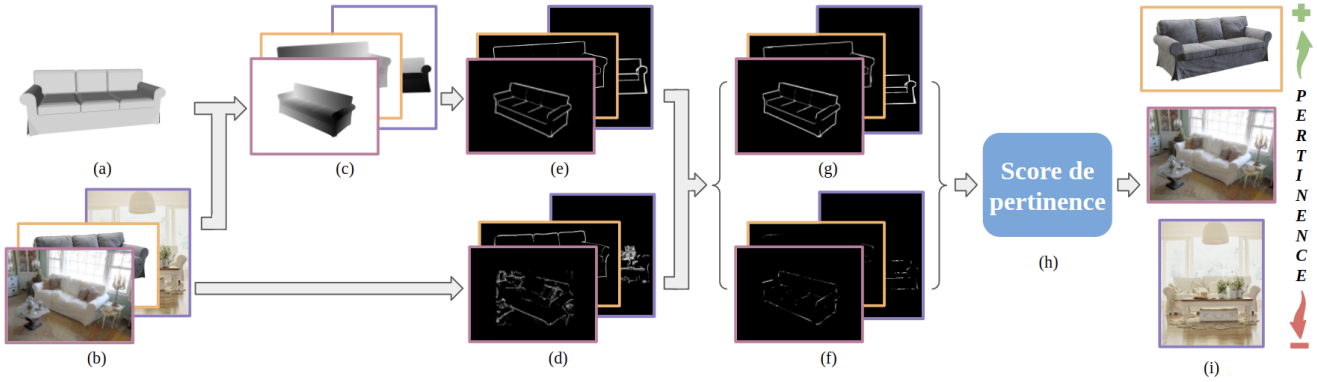

FIGURE 3 – Classement des images en fonction du *score de pertinence* : (a) le modèle 3D du l'objet d'étude et (b) l'ensemble des images contenant le même objet d'étude. Grâce à l'estimation de la pose, pour chaque image, nous calculons la carte de profondeur correspondante dans (c). Ensuite, nous estimons les cartes de saillance curviligne multi-échelle, MCS, associées à chaque image dans (d), ainsi que les cartes de saillance curviligne, CS, associées à chaque carte de profondeur, dans (e). Enfin, nous prenons l'intersection (f) et l'union (g) entre (d) et (e). Nous ne gardons que les points saillants appartenant à l'objet. Le *score de pertinence* (h) de chaque image conduit au classement des images dans (i). Dans la section 5, nous détaillons le calcul du *score de pertinence*.

sphère) pour générer une collection de cartes de profondeur [31]. Dans notre cas, comme la pose est donnée, nous générons uniquement les cartes de profondeur associées à la pose de chaque image. Ainsi, nous calculons la carte de saillance curviligne CS, cf. figure 3.(e). Les détails concernant le calcul de CS sont donnés dans l'annexe.

Les images originales, contrairement aux cartes de profondeur générées, contiennent des formes et des textures différentes. Par conséquent, comme préconisé dans [16], nous appliquons la saillance curviligne multi-échelle, MCS. Une pyramide gaussienne multi-échelle des images et les cartes de saillance curviligne associées sont calculées. Nous conservons les points d'intérêt qui apparaissent dans des échelles consécutives, cf. figure 3.(d). Les détails sur le calcul de MCS sont donnés dans l'annexe. Ainsi, pour chaque image, nous disposons de deux cartes de saillance curviligne, une issue de la carte de profondeur, et une autre, multi-échelle, relative à l'image.

La carte de profondeur fournit des informations caractéristiques relatives à la géométrie de l'objet alors que les points saillants de l'image peuvent être dus à la texture ou au contexte. Afin de ne garder que les points saillants appartenant à l'objet (et non à la texture), nous calculons l'intersection et l'union entre ces deux dernières cartes, cf. figure 3.(f) et (g). En effet, en prenant l'intersection entre la carte de saillance curviligne et la carte de saillance curviligne multi-échelle, nous souhaitons sélectionner uniquement les points saillants appartenant à l'objet. De manière complémentaire, l'union doit nous permettre de mettre en avant les points qui correspondent à un autre objet dans la scène. Elle permet également de pénaliser la localisation grossière des points saillants dans les images, c'est-à-dire les effets de flou sur la détection. Enfin, nous estimons deux nouvelles cartes s'appuyant sur un filtrage des cartes d' intersection et d'union, afin de limiter l'influence de la texture, c'est-à-dire l'utilisation de points détectés comme saillants alors qu'il s'agit d'une texture. Les détails sur le calcul de ces cartes filtrées sont donnés dans l'annexe.

La section suivante détaille la manière dont le *score de pertinence* est calculé à partir des cartes filtrées et de la boîte englobante des objets.

5 Score de pertinence

Les photographes disposent d'un ensemble de principes, et non de règles fixes, pour obtenir des photos de bonne qualité. Ces principes ont été utilisés pour détecter les zones d'attention dans les images et obtenir des cartes de saillance [32]. Par conséquent, le *score de pertinence* déterministe dépend de trois concepts inspirés, dans une certaine mesure, des pratiques de la photographie pour mettre en valeur l'objet capturé.

Informations caractéristiques ψ : Lors de la prise de vue, il est souvent recommandé de privilégier une vue qui montre le plus de détails possibles sur l'objet, tout en évitant les éléments de la scène qui peuvent attirer le regard et interférer avec l'attention portée au sujet principal de la photographie. L'objectif du premier terme, nommé « Informations caractéristiques », est d'appliquer ce concept en déterminant si l'environnement et la pose d'un objet dans une image donnée permettent d'extraire le plus d'informations caractéristiques possibles sur cet objet, voir les figures 2.(a) et (b). De plus, pour respecter cette règle photographique, l'objet d'étude doit être visible dans son intégralité, c'est-à-dire que l'objet ne doit être ni recadré ni occulté, figure 2.(d). Nous quantifions ces aspects en mesurant le nombre de points saillants visibles dans l'image : plus le nombre de points saillants est important, plus l'image est avantageuse. Plus précisément, nous pre-

nons le rapport entre le nombre de points saillants présents dans la carte d'intersection et le nombre de points saillants appartenant à la carte d'union, soit en calculant :

$$\psi = \frac{\#Intersection}{\#Union}, \tag{1}$$

où *Intersection* (respectivement *Union*) est l'ensemble des points saillants présents dans les deux (respectivement au moins une des deux) cartes CS et MCS.

Dominance β : En photographie, selon la règle des tiers, l'image est divisée par deux lignes verticales et horizontales et les objets d'intérêt doivent être situés soit sur ces lignes, soit sur leurs intersections, soit sur les diagonales de l'image. Un objet peut montrer beaucoup de détails mais être insignifiant dans une image très grande, voir les figures 2.(e) et (f). Une autre règle est d'utiliser un cadrage serré, ce qui signifie que l'objet est dominant dans l'image. Notre deuxième critère, nommé « Dominance », prend en compte cet aspect en calculant la taille de l'objet par rapport au contexte de l'image. Ainsi, nous avons introduit le paramètre de dominance, β, correspondant au rapport entre la taille de l'objet par rapport à la taille totale de l'image, soit plus précisément, le rapport des diagonales de sa boîte englobante et de l'image (de largeur et de hauteur respectives (w,h) et (W,H)) :

$$\beta = \frac{\sqrt{w^2 + h^2}}{\sqrt{W^2 + H^2}}. \tag{2}$$

Nous avons choisi une mesure linéaire car nos informations caractéristiques suivent des contours, c'est-à-dire des motifs linéaires, et varient donc linéairement.

Taille de l'objet γ : Naturellement, en photographie, on attend la meilleure résolution possible et notre dernier critère porte sur cet aspect. Ainsi, pour compléter, nous prenons en considération la taille réelle de l'objet, au travers du terme nommé « Taille de l'objet ». En effet, un objet, vu dans une image, peut avoir une pose pertinente, mais être petit. De plus, pour une même dominance et orientation, la taille de l'objet, et par conséquent la résolution de l'image, a une influence sur le nombre de points saillants détectés et cela peut avoir un impact sur la qualité de l'image. Plus la résolution est grande, plus il y a de points saillants disponibles. De la même façon que pour β, nous considérons une mesure linéaire pour l'objet et γ correspond à la longueur de la diagonale de sa boîte englobante :

$$\gamma = \sqrt{w^2 + h^2}. \tag{3}$$

Score de pertinence proposé : Notre score final correspond ainsi au produit de ces trois paramètres normalisés :

$$SP = \psi \times \beta \times \gamma. \tag{4}$$

Les détails concernant la normalisation des paramètres sont donnés dans l'annexe.

L'étude indépendante et conjointe des différents termes a été réalisée dans notre précédente publication. En effet, dans cette dernière [33], la favorabilité de la position d'un objet, relative à la dominance, est étudiée (cela signifie que l'on fait varier la position en fixant l'orientation). De plus, la pertinence de l'orientation, relative à la quantité d'information caractéristique visible, est évaluée séparément en fixant la position. Les résultats obtenus ont démontré l'intérêt de ces deux aspects.

6 Score de confiance

Certaines études ont montré que les réseaux de neurones sont capables d'imiter l'interprétation humaine. Par exemple, la similarité de forme et la similarité sémantique peuvent être évaluées par des réseaux de neurones de la même manière que par des personnes [34]. L'article [35] est une première étape importante dans la comparaison entre le comportement humain et celui des réseaux de neurones. Ils définissent la notion de typicalité, c'est-à-dire la caractéristique d'un objet qui est la plus remarquée par les humains. Ils montrent qu'un réseau de neurones peut évaluer la typicalité d'une image de manière similaire à celle attribuée par les humains. Cependant, les auteurs [36] soulignent un biais de la perception humaine : les humains catégorisent parfois les images en fonction de leur similarité avec un modèle « idéal ». Les réseaux peuvent être utilisés pour imiter le comportement humain dans des tâches de perception comme la prédiction de la capacité à mémoriser des objets dans les images [37]. La mémorisation est liée à un score d'importance attribué à chaque région segmentée de la scène. Ces travaux cités soulignent la capacité des réseaux à imiter la perception humaine et nous proposons donc de comparer notre approche déterministe avec un score alternatif utilisant les réponses des réseaux de neurones.

L'étude de la mise en valeur d'un objet donné dans une image peut être considérée comme un problème de classification ou de régression. Il est donc légitime de regarder le comportement des réseaux de neurones permettant de résoudre ces problèmes par rapport aux objectifs que nous visons et notamment d'étudier si leurs sorties peuvent apporter une réponse alternative à l'approche déterministe que nous avons développée. Notre ligne directrice est la suivante : un objet correctement mis en valeur doit être facilement détectable ainsi que reconnaissable sans ambiguïté par un réseau de neurones dédié à la détection ou à la classification d'objets. Dans notre étape de validation, nous avons choisi de considérer quatre réseaux de neurones : un réseau traditionnel de détection d'objets *Faster R-CNN* [28] et le réseau le plus reconnu dans la littérature, à savoir *YOLOv5* [38], la dernière version de *YOLO* [30], tous deux pré-entraînés sur la base de données *COCO* [39]. Pour les réseaux de classification, nous utilisons également les plus performants dans la littérature, *EfficientNet* [25] et *EfficientNet2* [40], pré-entraînés sur la base de données *Imagenet*.

7 Construction de classements de référence

L'approche proposée permet de classer un ensemble d'images en fonction de leur capacité à mettre en valeur un objet donné. À notre connaissance, il n'existe pas de classement de référence d'images en fonction de la mise en valeur d'un objet donné. Nous avons donc choisi de générer des ensembles d'images contenant des dégradations que nous contrôlons afin de construire une classement de référence associé. Ce classement nous permet donc d'évaluer mais également de comparer nos résultats de classement déterministe avec les classements obtenus en utilisant des réseaux de neurones. La suite de cette section nous permet de fournir les détails pour construire les images dégradées.

Nous créons un ensemble d'images pour chaque objet étudié. Pour ce faire, nous choisissons une image initiale à laquelle nous appliquons successivement trois différents type de dégradations, cf. figure 4 :

- Augmentation de la taille de l'arrière-plan et donc de l'image sans modifier la taille de l'objet, cf. figure 4.(b) ;
- Ajout d'occultations, en insérant d'autres objets dans l'image, cf. figure 4.(c) ;
- Diminution de la taille de l'image, ce qui induit une diminution de la résolution et donc de l'objet d'intérêt qu'elle contient, cf. figure 4.(d).

Nous itérons ces trois dégradations, de manière séquentielle et dans cet ordre, tant que la résolution de l'image reste acceptable. Une fois que ces dégradations ont été effectuées, nous obtenons un ensemble d'images ordonnées de la moins dégradée (image initiale) à la plus dégradée. En appliquant séquentiellement les dégradations, l'ordre des images est objectivement déterminé.

8 Résultats obtenus pour les classements de référence

L'approche proposée nous permet de classer les images du score le plus élevé au score le plus bas, en utilisant le *score de pertinence* ou le score de confiance. Afin d'évaluer l'ordre proposé, nous utilisons la corrélation d'ordre de rang de Spearman, *Spearman rank order correlation coefficient*, SROCC, comme dans [35]. Cette corrélation utilise le coefficient de corrélation linéaire de Pearson, *Pearson Linear Correlation Coefficient*, PLCC, avec les variables de rang. La mesure SROCC est définie comme suit :

$$SROCC(X,Y) = \frac{cov(R(X),R(Y))}{\rho_{R(X)} \cdot \rho_{R(Y)}} \qquad (5)$$

avec :

- $R(X)$ variable de rang,
- $cov(R(X),R(Y))$ covariance de $R(X)$ and $R(Y)$,
- $\rho_{R(X)}$ et $\rho_{R(Y)}$ écart-type de $R(X)$ and $R(Y)$.

Les valeurs de SROCC varient entre -1 et 1.

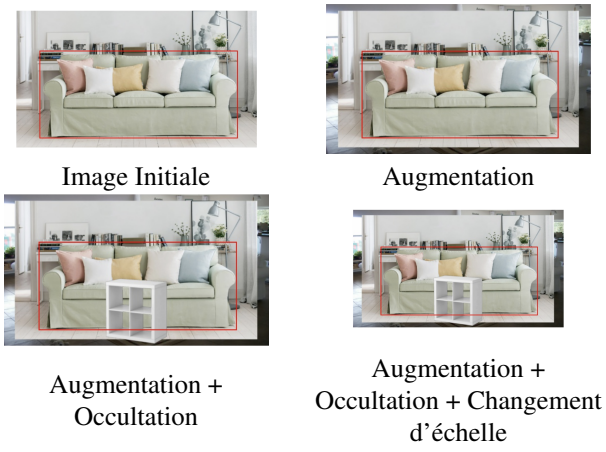

FIGURE 4 – Exemple des trois dégradations considérées. Dans cette figure nous appliquons successivement les trois types de dégradations et cela nous permet d'avoir une classification objective de référence de (a) à (d).

Étant donné le classement objectif, nous calculons la corrélation entre l'ordre précédemment établi comme référence, et les ordres estimés en s'appuyant sur le *score de pertinence* et le score de confiance. Il est important de noter que nous ne conservons que les classements dont les corrélations sont significatives, c'est-à-dire les classements qui ont obtenu une corrélation avec une valeur de probabilité (p) ou *p-value* inférieure à un seuil α. Dans la suite, nous avons choisi de considérer deux valeurs habituellement utilisées dans la littérature pour le seuil α : 0.05 et 0.01.

Dans la Table 1, pour chaque méthode, nous avons calculé les corrélations moyennes sur les mêmes 40 classements significatifs avec $\alpha = 0.05$. Parmi les réseaux, *EfficientNet* est le plus performant, nous conservons donc ses résultats pour une comparaison ultérieure. Le réseau *Faster R-CNN* a été étudié mais les résultats n'étaient pas assez compétitifs pour être conservés. Pour *YOLOv5*, le score de confiance est un produit entre la probabilité d'obtenir la classe dans la boîte englobante estimée et un *score de pertinence* de la boîte englobante déterminée, c'est-à-dire à quel point elle respecte les contours de l'objet. Nous avons essayé d'extraire seulement la probabilité de la classe comme score de sortie, mais les résultats ne sont pas meilleurs, donc nous avons gardé le score de confiance original dans les résultats que nous présentons. Selon la Table 1, notre méthode est celle qui présente la meilleure corrélation moyenne.

Méthodes	Proposée	*YOLOv5*	*EfficientNet*	*EfficientNet2*
SROCC	**0.902**	0.545	0.836	0.661

TABLE 1 – Corrélation moyenne sur 40 classements significatifs. En gras, nous indiquons le meilleur résultat obtenu.

(a) n°1 (b) n°2 (c) n°3 (d) n°4

(e) n°1 (f) n°2 (g) n°3 (h) n°4

FIGURE 5 – Classements automatiques obtenus avec la méthode déterministe proposée : du *score de pertinence* le plus élevé (n°1) au *score de pertinence* le plus bas (n°4).

	$\alpha = 0.05$		$\alpha = 0.01$	
Méthodes	Proposée	*EfficientNet*	Nous	*EfficientNet*
SROCC	**0.898**	0.795	**0.921**	0.823

TABLE 2 – Corrélation moyenne. Respectivement, avec $\alpha = 0,05$ (respectivement $\alpha = 0,01$), nous avons 116 (respectivement 84) classements significatifs en commun. Nous notons en gras les meilleurs résultats.

La Table 2 compare les deux meilleurs classements obtenus avec *EfficientNet* et notre méthode déterministe sur les classements significatifs pour les deux méthodes. Pour le seuil de la *p-value*, deux valeurs $\alpha \in \{0,05, 0,01\}$ ont été testées. Les résultats montrent que notre méthode est plus performante que *EfficientNet*.

Nous remarquons que les réseaux *YOLOv5* et *EfficientNet2* ne sont pas performants. Toutefois, il est important de noter que, d'une part, les résultats de ces réseaux dépendent du jeu de données sur lequel ils ont été entraînés, et, d'autre part, ils n'ont pas été proposés pour répondre à la tâche que nous étudions. En particulier, les réseaux ne sont pas toujours affectés par le changement de taille ou de dominance à cause de l'augmentation des données qui peut être effectuée pendant la phase d'apprentissage. Ainsi, ils sont robustes à ces deux types de dégradations (augmentation et mise à l'échelle). En revanche, l'ajout d'occultations les affecte, comme déjà mentionné dans [41]. Ainsi, leurs scores de confiance ne seront pas affectés par l'augmentation ou les changements d'échelle mais ils varieront en fonction de la présence d'occultations ou non. Pour confirmer ce comportement, nous avons testé l'estimation d'un classement pour un ensemble d'images ayant subi uniquement des ajouts d'occultations. De la même façon que pour les tests précdéents, seuls les classements avec une corrélation significative sont conservés. De plus, nous avons choisi $\alpha = 0.05$. Sur cet exemple, la Table 3 montre que *YOLOv5* et *EfficientNet2* obtiennent de meilleurs résultats que ceux de la Table 1. Cela confirme le comportement attendu.

Méthodes	*YOLOv5*	*EfficientNet*	*EfficientNet2*
SROCC	0.826	**0.847**	0.817

TABLE 3 – Corrélation moyenne sur 19 classements significatifs qui ne contiennent que des occultations.

Ces résultats montrent que *YOLOv5* et *EfficientNet2* ne prennent pas en compte le changement d'échelle, c'est-à-dire que les scores de confiance ne sont pas affectés par le changement d'échelle, mais sont impactés par les occultations. Or, pour choisir des images qui mettent en valeur un objet, il est essentiel de prendre en compte le changement d'échelle et de ne pas être affecté par les occultations. Ainsi, nous avons montré que *YOLOv5* et *EfficientNet2* restent tous deux moins efficaces que *EfficientNet*. En conclusion, notre méthode déterministe et *EfficientNet* sont les seules capables de prendre en compte les trois types de dégradation.

9 Résultats sur des données réelles

Base de données : Nous avons besoin d'images qui contiennent un objet à étudier, ainsi que le modèle 3D

de cet objet. Pour avoir une telle correspondance, nous avons choisi la base de données Pix3D [42] qui contient des images de différents canapés et les modèles 3D correspondants. Pour chaque image, la pose 3D de l'objet est connue et permet d'assurer un alignement 2D-3D. Nous avons testé notre approche sur 9 modèles de la catégorie *Sofa*, soit 1048 images.

Classements : Les résultats obtenus permettent de valider visuellement l'approche proposée, comme le montrent les exemples de classement présentés dans les figures 1 et 5. Nous remarquons que les dernières images mettent le moins en valeur l'objet car il est tronqué ou mal positionné. En revanche, les images classées en premier correspondent soit à des images où l'objet est dominant dans la scène, soit à des images où il est acquis dans un environnement neutre ou il y a peu d'occultations. Pour pouvoir comparer nos résultats sur données réelles à ceux obtenus en utilisant de l'apprentissage, nous avons donné les images du premier classement aux différents réseaux de neurones étudiés. Les classements ainsi obtenu sont présentés en annexe dans la figure 7.

10 Conclusion

Nous avons proposé une méthode permettant de mesurer la qualité subjective d'une image en quantifiant sa capacité à mettre en valeur un objet donné. En utilisant cette quantification représentée par un *score de pertinence*, nous sommes en mesure de proposer des classements d'images et ainsi d'automatiser le choix de l'image la plus avantageuse. Aucun entraînement sur un grand jeu de données n'est nécessaire. Les réseaux de neurones ont montré des performances plus faibles, à l'exception du réseau *EfficientNet* qui rivalise avec notre méthode.
Une première perspective de ce travail est de considérer des primitives saillantes non plus dans une carte de profondeur (projection 2D du modèle 3D suivant différents points de vue) mais directement dans le modèle 3D. À plus longs termes, nous souhaitons généraliser ce travail à des données temporelles.

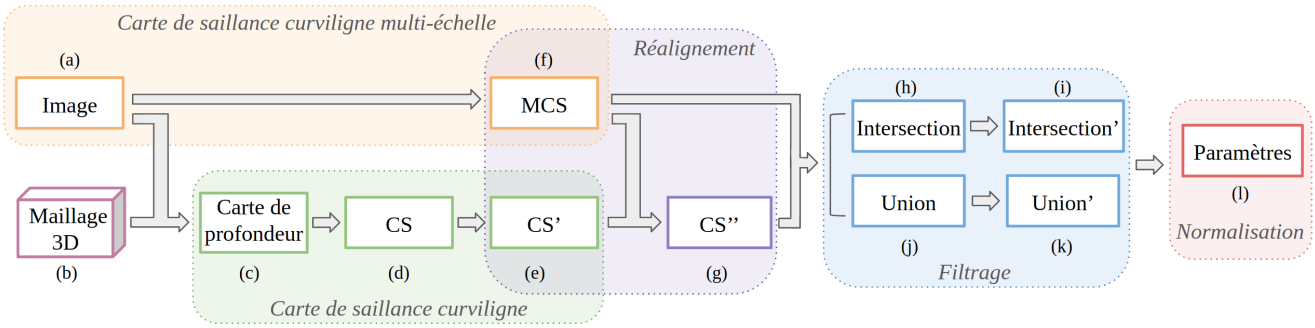

FIGURE 6 – Méthode proposée s'appuyant sur les *scores de pertinence* – En entrée, nous avons une image contenant un objet, en (a), et le modèle 3D de cet objet, en (b). Pour pouvoir les comparer, nous calculons d'abord la carte de profondeur, en (c), à partir du modèle 3D et de la pose donnée. Nous estimons ensuite la carte de saillance curviligne multi-échelle, MCS, associée à l'image, en (f), et la carte de saillance curviligne, CS, associée à la carte de profondeur, en (d). Les valeurs inférieures et non significatives de CS sont filtrées en (e). Avant de comparer les cartes MCS et CS, un ré-alignement global est calculé en (g), et finalement, l'intersection (h) et l'union (j) sont estimées. Les versions filtrées (i) et (k) sont calculées pour générer les paramètres du *score de pertinence*, qui sont ensuite normalisés en (l).

Annexes

Carte de saillance curviligne : Etant donné une carte de profondeur, figure 7.(c), nous calculons sa carte de saillance curviligne (CS), figure 7.(d), en appliquant la formule suivante à tous les points p de la carte de profondeur :

$$CS(p) = \lambda_1(p) - \lambda_2(p) \tag{6}$$

avec $\lambda_2(p) \leq \lambda_1(p)$.
Pour un point p donné, les valeurs $\lambda_2(p)$ et $\lambda_1(p)$ sont les courbures principales de la carte de profondeur au point p. Elles sont facilement calculées comme les valeurs propres de la matrice hessienne $H(p)$ de la surface représentée par la carte de profondeur, qui est une matrice diagonale dans ce cadre fonctionnel [16]. Pour calculer la carte CS, nous lissons d'abord la surface associée en appliquant un filtre gaussien. Nous avons choisi empiriquement $\sigma = 1.4$ qui offre un compromis entre le lissage et le floutage. Ensuite, nous filtrons le bruit dû aux valeurs CS trop proches de 0 (figure 7.(e)) , en éliminant tous les points de la carte CS qui ont une valeur de saillance curviligne inférieure à 1% de la valeur de saillance curviligne maximale :

$$\forall p \in CS, \begin{cases} p \text{ est gardé} & \text{si} \quad CS(p) \geq 0.01 \times CS_{max} \\ p \text{ est supprimé} & \text{sinon.} \end{cases} \tag{7}$$

avec CS_{max} la valeur maximale de CS.

Carte de saillance de curviligne multi-échelle : Pour effectuer le classement des images de la moins favorable à la plus favorable à l'objet, nous avons utilisé la saillance curviligne multi-échelle (MCS) sur l'image 2D (figure 7.(f)). Plus précisément, nous construisons, pour chaque image, une pyramide gaussienne avec $nb\_echelle = 4$ échelles, filtrée par :

$$\forall i \in [1, nb\_echelle],\ \sigma_i = \sigma k^{i-1} \tag{8}$$

avec $k = 2^{\left(\frac{1}{nb\_echelle}\right)}$ and $\sigma = 1.4$.
De plus, pour réduire l'impact de la texture qui provoque des valeurs CS très faibles par rapport celles induites par les changements de géométrie et pour limiter le temps de calcul, à chaque échelle, seuls les points ayant les valeurs de saillance curviligne les plus élevées sont conservés. De plus, autant de points saillants que dans la carte CS, avec une marge de 10% d'erreur, sont sélectionnés. Enfin, parmi ces derniers, seuls les points qui sont présents dans au moins $N = 3$ échelles consécutives sont sauvegardés.

Réalignement : La base de données Pix3D fournit un ensemble d'images pour chaque modèle de mobilier 3D. La pose et le calibrage de chaque objet dans chaque image sont connus et nous permet de générer les cartes de profondeur. Cependant, la correspondance entre l'image et la carte de profondeur n'est pas parfaite. Pour corriger l'erreur de correspondance, un réalignement de la carte CS par rapport à la carte MCS est estimé afin de maximiser le nombre de points saillants présents dans l'intersection binaire (cf. figure 7.(g). Plus précisément, nous étudions l'intersection entre la carte CS et la carte MCS translatée par un vecteur $(i, j), \forall i, j \in [-5, 5]$.

Filtrage : Pendant le filtrage de la carte d'intersection (figure 7.(i)), nous faisons l'hypothèse que si deux points sont homologues, alors leurs voisinages doivent être similaires, c'est-à-dire contenir la même distribution de points d'intérêt. Ainsi, nous comparons les deux voisinages des deux points correspondants entre l'image et la carte de profondeur. Dans notre étude, deux points sont considérés comme appariés lorsqu'ils sont détectés à la même position, respectivement dans la carte CS et la carte MCS. Malgré l'étape de réalignement, on peut supposer qu'il existe parfois un décalage entre les positions des deux points, ce qui correspond à une transformation rigide. C'est pourquoi une

(a) n°1 (b) n°2 (c) n°3 (d) n°4

(e) n°1 (f) n°2 (g) n°3 (h) n°4

(i) n°1 (j) n°2 (k) n°3 (l) n°4

FIGURE 7 – Classements automatiques obtenus avec les méthodes d'apprentissage étudiées (*YOLOv5* en 1er ligne, *EfficientNet* en 2e ligne et *EfficientNet2* en 3e ligne)
: du *score de confiance* le plus élevé (n°1) au *score de pertinence* le plus bas (n°4).

fenêtre glissante a été considérée. Cependant, l'ajout de cette fenêtre glissante n'améliore pas la qualité des résultats.

Il existe dans la littérature plusieurs métriques capables de mesurer le degré de similarité entre deux voisinages et nous avons choisi la distance de Hausdorff. Nous désignons p_i et p_c deux points détectés respectivement dans la carte MCS (par rapport à l'image) et dans la carte CS (liée à la carte de profondeur associée). Nous désignons par N la longueur de côté du voisinage carré pris en compte, $\mathbf{V}_i$ le voisinage de p_i dans la carte MCS et $\mathbf{V}_c$ pour le voisinage de p_c dans la carte CS. La valeur de N dépend de la taille de la boîte englobante de l'objet étudié. Les deux voisinages $\mathbf{V}_c$ et $\mathbf{V}_i$ sont préalablement binarisés :

$$\forall k \in \{i, c\},\ \forall\, lig, col \in [1, N]^2,$$

$$\mathbf{V}_k[lig, col] = \begin{cases} 1 & \text{si} \quad \mathbf{V}_k[lig, col] \neq 0 \text{ (point saillant)} \\ 0 & \text{sinon.} \end{cases} \tag{9}$$

Et la distance est donnée par :

$$d(\mathbf{V}_c, \mathbf{V}_i) = \max \{\sup_{p_c \in \mathbf{V}_c} \inf_{q_i \in \mathbf{V}_i} \delta(q_i, p_c), \sup_{q_i \in \mathbf{V}_i} \inf_{p_c \in \mathbf{V}_c} \delta(q_i, p_c)\} \tag{10}$$

Un point saillant est conservé si son voisinage et celui de son point homologue présentent une similarité d'au moins 70% selon la distance de Hausdorff.

Le problème est similaire lors de la création de la carte d'union (cf. figure 7.(j)). En effet, un processus de filtrage est nécessaire lorsqu'un point saillant est détecté dans l'image et non dans la carte de profondeur. L'objectif est de déterminer si la différence est due à la texture, ou à un décalage local. Comme pour la carte d'intersection, nous mesurons la distance de Hausdorff entre les deux voisinages.

Normalisation : À la fin du pipeline, nous pouvons extraire les trois paramètres nécessaires au calcul du score de mise en évidence (cf. figure 7.(k)). Pour s'assurer que les trois paramètres ont la même importance, nous les avons normalisés (les valeurs sont comprises entre $0, 1$ et 1).

Références

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet : A large-scale hierarchical image database," in *IEEE Conference on Computer Vision Pattern Recognition, CVPR*, 2009.

[2] T. Malon, P. Guyot, G. Roman-Jimenez, S. Chambon, V. Charvillat, A. Crouzil, A. Péninou, J. Pinquier, F. Sèdes, and C. Sénac, "Toulouse campus surveillance dataset : scenarios, soundtracks, synchronized videos with overlapping and disjoint views," in *ACM Multimedia Systems Conference, MMSys*, 2018.

[3] Y. Xiang, R. Mottaghi, and S. Savarese, "Beyond PASCAL : A Benchmark for 3D Object Detection in the Wild," in *IEEE Winter Conference on Applications of Computer Vision, WACV*, 2014.

[4] T. Collins, D. Pizarro, S. Gasparini, N. Bourdel, P. Chauvet, M. Canis, L. Calvet, and A. Bartoli, "Augmented reality guided laparoscopic surgery of the uterus," *IEEE Transactions on Medical Imaging, TMI*, vol. 40, no. 1, pp. 371–380, 2020.

[5] A. Biglia and P. Belleflamme, "Analyse prospective sur l'implémentation de la voiture autonome : impact sur l'industrie automobile et le citoyen," Université Catholique de Louvain, Belgique, Mémoire de master, 2015.

[6] M. Ben Amor, A. Samet, F. Kammoun, and N. Masmoudi, "Exploitation des caractéristiques du système visuel humain dans les métriques de qualité," in *Applications Médicales de l'Informatique : Nouvelles Approches*, 2010.

[7] M. Herbin, M. Chambah, E. Zagrouba, and S. Ouni, "Vers une métrique de description objective d'une sensation subjective," *Revue Africaine de la Recherche en Informatique et Mathématiques Appliquées*, 2009.

[8] T.-Y. Pan, Y.-Z. Dai, W.-L. Tsai, and M.-C. Hu, "Deep model style : Cross-class style compatibility for 3d furniture within a scene," in *IEEE International Conference on Big Data (Big Data)*, 2017.

[9] C. Harris and M. Stephens, "A combined corner and edge detector," in *Alvey Vision Conference, AVC*, 1988.

[10] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *European Conference on Computer Vision*, 2006.

[11] S. Smith and J. Brady, "SUSAN – A New Approach to Low Level Image Processing," *International Journal on Computer Vision, IJCV*, vol. 23, no. 1, pp. 45–78, 1997.

[12] T. Tuytelaars and L. Van Gool, "Matching Widely Separated Views Based on Affine Invariant Regions," *International Journal on Computer Vision, IJCV*, vol. 59, pp. 61–85, 2004.

[13] P. Fischer and T. Brox, "Image descriptors based on curvature histograms," in *German Conference on Pattern Recognition*, 2014.

[14] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal on Computer Vision, IJCV*, 2004.

[15] G. Flitton, T. Breckon, and N. Megherbi Bouallagu, "Object recognition using 3d sift in complex ct optvolumes," in *British Machine Vision Conference, BMVC*, 2010.

[16] H. A. Rashwan, S. Chambon, P. Gurdjos, G. Morin, and V. Charvillat, "Using Curvilinear Features in Focus for Registering a Single Image to a 3D Object," *IEEE Transactions on Image Processing, TIP*, 2019.

[17] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[18] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *European Conference on Computer Vision*. Springer, 2016.

[19] M. Xu, J. Zhong, Y. Ren, S. Liu, and G. Li, "Context-aware attention network for predicting image aesthetic subjectivity," in *ACM International Conference on Multimedia*, 2020.

[20] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Wang, "Rapid : Rating pictorial aesthetics using deep learning," in *ACM International Conference on Multimedia*, 2014.

[21] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *IEEE Conference on Computer Vision Pattern Recognition, CVPR*, 2017, pp. 3156–3164.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR (arXiv)*, 2014.

[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conference on Computer Vision Pattern Recognition, CVPR*, 2015.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision Pattern Recognition, CVPR*, 2016.

[25] M. Tan and Q. Le, "Efficientnet : Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*, 2019.

[26] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conference on Computer Vision Pattern Recognition, CVPR*, 2014.

[27] R. Girshick, "Fast r-cnn," in *IEEE Conference on Computer Vision Pattern Recognition, CVPR*, 2015.

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

[29] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. Berg, "Ssd : Single shot multibox detector," *Lecture Notes in Computer Science*, 2016.

[30] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once : Unified, real-time object detection," in *IEEE Conference on Computer Vision Pattern Recognition, CVPR*, 2016.

[31] C. Bongsoo Choy, M. Stark, and S. Corbett-Davies, S.and Savarese, "Enriching object detection with 2d-3d registration and continuous viewpoint estimation," in *IEEE Conference on Computer Vision Pattern Recognition, CVPR*, 2015, pp. 2512–2520.

[32] E. Kozegar, "Rule of photography in image saliency detection," 2016.

[33] M. Pelissier Combescure, G. Morin, and S. Chambon, "Extraction et comparaison d'information saillante : Pose favorable et image 2d révélatrice d'un objet 3d," in *ORASIS*, 2021.

[34] J. Kubilius, S. Bracci, and H. Op de Beeck, "Deep neural networks as a computational model for human shape sensitivity," *PLoS computational biology*, vol. 12, no. 4, 2016.

[35] B. Lake, W. Zaremba, R. Fergus, and T. Gureckis, "Deep neural networks predict category typicality ratings for images," in *Cognitive Science, CogSci*, 2015.

[36] L. Barsalou, "Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories." *Journal of experimental psychology : learning, memory, and cognition*, vol. 11, no. 4, 1985.

[37] R. Dubey, J. Peterson, A. Khosla, M. Yang, and B. Ghanem, "What makes an object memorable ?" in *IEEE Conference on Computer Vision Pattern Recognition, CVPR*, 2015.

[38] G. Jocher, "Yolov5, https ://github.com/ultralytics/yolov5.git," 2020.

[39] T. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. Zitnick, and P. Dollár, "Microsoft COCO : Common Objects in Context," *CoRR (arXiv)*, vol. abs/1405.0312, 2014.

[40] M. Tan and Q. Le, "Efficientnetv2 : Smaller models and faster training. arxiv 2021," *CoRR (arXiv)*.

[41] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. Wichmann, and W. Brendel, "Imagenet-trained cnns are biased towards texture ; increasing shape bias improves accuracy and robustness," *CoRR (arXiv)*, 2018.

[42] X. Sun, J. Wu, X. Zhang, Z. Zhang, C. Zhang, T. Xue, J. B. Tenenbaum, and W. Freeman, "Pix3D : Dataset and methods for single-image 3D shape modeling," in *IEEE Conference on Computer Vision Pattern Recognition, CVPR*, 2018.