Suppression de spécularités par réseau adverse multi-classes

J. Lin[1,2] M. E. A. Seddik[1] M. Tamaazousti[1] Y. Tamaazousti[1] A. Bartoli[2]

[1] CEA, LIST, Point Courrier 94, Gif-sur-Yvette, F-91191, France
{john.lin, MohamedElAmine.Seddik, mohamed.tamaazousti}@cea.fr
youssef.tamaazousti@gmail.com

[2] Institut Pascal - UMR 6602 - CNRS/UCA/CHU, Clermont-Ferrand, France
adrien.bartoli@gmail.com

Résumé

La séparation des composantes de réflexion diffuse et spéculaire est un problème complexe et ouvert dû à l'ambiguïté du processus de formation d'images. Nous proposons une nouvelle approche de suppression de spécularités, sous la forme d'un réseau convolutif qui prend une image en entrée et génère en sortie sa partie diffuse. Il s'agit d'un réseau génératif adverse, où la fonction de coût adverse est combinée à une fonction de coût perceptuelle, qui est un terme d'attache aux données. En contraste avec les réseaux adverses classiques, le discriminateur est multi-classes et non binaire, ce qui lui permet de se concentrer sur des caractéristiques plus pertinentes des images. De manière plus formelle, cela apporte deux termes de gradient supplémentaires pour trouver la distribution du domaine diffus. Nous entraînons notre modèle sur une base de données synthétiques, que nous avons pensées spécifiquement pour la tâche demandée. Nous montrons enfin que notre méthode opère de manière plus consistente sur une plus grande diversité de scènes que l'état de l'art.

Mots Clef

Deep learning, spécularité, réseau de neurones convolutif.

Abstract

We propose a novel learning approach, in the form of a fully-convolutional neural network, which automatically and consistently removes specular highlights from a single image by generating its diffuse component. To train the generative network, we define an adversarial loss on a discriminative network as in the GAN framework and combined it with a content loss. In contrast to existing GAN approaches, we implemented the discriminator to be a multi-class classifier instead of a binary one, to find more constraining features. This helps the network pinpoint the diffuse manifold by providing two more gradient terms. We also rendered a synthetic dataset designed to help the network generalize well. We show that our model performs well across various synthetic and real scenes and outperforms the state-of-the-art in consistency.

(a) Input (b) Diffuse (c) Specular

FIGURE 1 – Séparation spéculaire-diffus avec notre méthode. Notre modèle prend une image en entrée (a) et génère automatiquement sa partie diffuse (b) sans a priori sur la scène. La partie spéculaire (c) est obtenue en soustrayant (b) à (a).

Keywords

Deep learning, specularity, convolutional neural network.

1 Introduction

L'apparence d'un objet dépend de la manière dont il reflète la lumière, les matériaux les plus communs ayant un comportement dichromatique : ils produisent deux types de réflexion à leur surface, à savoir diffuse et spéculaire. Le modèle dichromatique de Shafer [30] décrit la formation d'une image par la combinaison linéaire de ces deux termes. Contrairement aux réflexions diffuses qui sont uniformes dans toutes les directions, la composante spéculaire dépend de l'angle d'observation. Cette dernière est donc responsable de nombreuses difficultés rencontrées en problèmes de vision par ordinateur. C'est pourquoi de nombreux algorithmes assument que la surface des objets est purement diffuse et ne prennent pas en compte la réflexion spéculaire. C'est le cas du SLAM, de la segmentation d'image ou de la détection d'objets pour n'en citer que quelques-uns. Cependant, cela se traduit par une défaillance des méthodes lorsque la surface de l'objet s'éloigne, parfois mêmes infimement, d'un diffuseur parfait.

Naturellement, diverses approches ont été proposées pour

résoudre ce problème de séparation spéculaire-diffus, dans l'idée d'être appliquées comme pré-traitement à d'autres algorithmes. Une telle approche est également utile dans le domaine de l'infographie puisque les spécularités véhiculent de précieuses informations concernant la scène observée [9, 20]. Toutefois, le problème reste difficile et ouvert car il est mal posé à cause des nombreux paramètres et de l'ambiguïté du processus de formation d'images [1]. Bien que certains travaux [2, 31] montrent des résultats convainquants, ils manquent de souplesse dans leur utilisation car ils ont besoin d'équipements spéciaux [21, 25], d'un pré-traitement tel que de la segmentation de couleurs [3, 15, 30], ou encore de connaître des a prioris sur la scène comme la couleur de l'illuminant [21, 34, 35].

En contraste avec les travaux de la littérature, nous proposons une approche par apprentissage pour surmonter les limites d'utilisation. L'idée est que le réseau estimera de lui-même les relations complexes entre une image et sa partie diffuse, à travers une analyse probabilistique sur un grand nombre de données. Une telle approche nous permet de nous affranchir de la nécessité d'expliciter nous-mêmes des caractéristiques et des a prioris, qui pourraient s'avérer ne pas être pertinents pour la diversité des scènes possibles [36]. Une difficulté rencontrée immédiatement est l'absence d'une base de données réelles adaptée au problème. Une telle base est impossible ou au moins extrêmement longue à produire. Nous entraînons donc notre réseau sur une base de données synthétiques, que nous générons spécifiquement pour surmonter des cas difficiles en les incluant directement dedans.

Pour aider à la bonne généralisation de la méthode, d'autant plus difficile avec des données synthétiques, nous entraînons notre modèle dans le cadre des réseaux génératifs adverses [8] que nous adaptons au problème de séparation des composantes de réflexion. Notre architecture est constituée de deux réseaux : un réseau de neurones convolutif qui supprime les spécularités à partir d'une seule image et un discriminateur spécifique à cette tâche de suppression. Ce dernier a pour rôle de déterminer si les spécularités ont été bien retirées. Contrairement à une architecture adverse classique, notre discriminateur est un classifieur multi-classes et non binaire. En augmentant le nombre de classes, nous aidons le réseau à identifier la distribution recherchée grâce à un gradient plus précis. Nous montrons dans §4 que notre méthode est plus précise et plus stable pour des images plus diverses que les approches existantes, à la fois qualitativement et quantitativement.

En résumé, cet article adresse les défis mentionnés plus haut et met en avant trois contributions :

— Une nouvelle méthode de séparation spéculaire-diffus à partir d'une seule image, sans a priori sur la scène et capable d'opérer sur diverses images.
— Une fontion de coût adverse multi-classes pour le problème de séparation.
— Une nouvelle base de données, pensée pour la suppression de spécularités.

2 État de l'art

Nous regroupons les méthodes existantes parmi trois catégories : les méthodes basées modèle à image unique et à images multiples et les méthodes basées apprentissage.

Méthodes à images multiples. Une différence entre la lumière spéculaire et la lumière diffuse est que la première est polarisée tandis que la dernière ne l'est pas. Sachant cette propriété, [37] utilise un polariseur afin de séparer ces composantes de réflexion dans des images en niveau de gris. [25] étend ces travaux en ajoutant une contrainte colorimétrique : la méthode évalue la couleur de l'illumination, qui est la même que celle des spécularités [30], et l'utilise pour contraindre la couleur diffuse. L'utilisation de polariseurs donne de bons résultats même pour une surface texturée mais rend la méthode peu pratique puisqu'il faut un équipement supplémentaire qui doit en plus être tourné dans la bonne orientation. [21] se sert de la dépendance des spécularités au point de vue. En effet, un point sur une surface peut appraître spéculaire dans une image et purement diffus dans une image d'un autre point de vue. En trouvant ces paires de points dans deux vues différentes, on peut supprimer les spécularités d'une image. De manière semblable, [22, 28] proposent une méthode à images multiples, mais avec une lumière mobile au lieu d'un point de vue changeant. [13] utilise la même propriété mais capture un champ lumineux en déplaçant une caméra autour de la surface et en prenant le flux vidéo en entrée. Ces méthodes nécessitent une géométrie connue de la surface ou considèrent des surfaces planes. Ces méthodes à images multiples parviennent à séparer les composantes de réflection, même pour des surfaces texturées car les opérations sont locales sans contrainte de pixels voisins. En revanche, elles sont peu pratiques, notamment pour une utilisation temps réel dans un flux vidéo. Nous nous concentrons donc sur des méthodes à image unique.

Méthodes à image unique. Ces méthodes opèrent généralement en analysant la couleur d'un objet sous les hypothèses du modèle dichromatique de Shafer [30]. Ce dernier a été le premier à proposer une méthode de séparation à partir d'une seule image couleur, en constatant que la distribution des pixels d'une surface uniforme forme un parallélogramme dans l'espace RGB. [15] étend ces travaux en observant que l'histogramme des chromaticités d'une surface uniforme forme un "T oblique" où les deux branches représentent les pixels purement diffus et les pixels spéculaires. La suppression des spécularités est ensuite réalisée en projetant les pixels spéculaires sur la branche diffuse. Cependant, en application réelle on ne peut pas toujours identifier ce "T" à cause du bruit, des interréflexions et des multiples sources de lumières. Cette méthode nécessite également une segmentation par couleur pour les objets de texture non uniforme, qui peut justement s'avérer difficile en présence de spécularités et pour des scènes complexes. [35] établit une méthode ne nécessitant pas une telle segmentation. Celle-ci génère une image purement diffuse

en décalant l'intensité et la chromaticité des pixels, tout en conservant leur teinte. Cette image, appelée "specular-free image", possède la même structure géométrique que l'image originale et ne contient aucune spécularité. Elle ne diffère que par la couleur des surfaces, celle-ci étant choisie arbitrairement. Par différentiation logarithmique des intensités entre cette "specular-free image" et l'image originale, on peut alors déterminer si les pixels de l'image originale contiennent des spécularités et ainsi les retirer en soustrayant une petite constante itérativement aux trois canaux RGB. Le concept de "specular-free image" a été largement utilisé dans la littérature, offrant parfois des variantes. Notamment, [32] propose une "modified specular-free image" plus robuste et de la même façon, estime l'image diffuse par optimisation itérative sur la différence entre l'image originale et la "modified specular-free image". Ces méthodes sont limitées par la nécessité de n'avoir qu'une couleur d'illumination, parfaitement blanche ou connue.

Méthodes basées apprentissage. Il y a peu de travaux traitant du problème de séparation des composantes de réflexion. [40] utilise la structure de translation de domaines, comme le cycleGAN [39], afin d'entraîner un réseau qui passe des images du domaine spéculaire au domaine diffus. Cependant, leur méthode est pensée pour la reconstruction 3D multi-vues. Elle prend donc plusieurs images en entrée et renvoie des images du même objet avec un matériau diffus blanc, ce peu importe la matériau originel. Cette tâche est plus simple car elle ne nécessite pas de décorréler l'apport de l'illumination et du matériau dans la couleur. [7] adopte aussi la structure du cycleGAN mais ne travaille que sur des images endoscopiques médicales, qui est un domaine assez restraint. En contraste avec les méthodes basées modèle, les méthodes par apprentissage ne reposent pas sur des forts a prioris tels que la sparsité de la réflectance dans la distribution RGB [27] ou des a prioris sur la géométrie et l'illumination [4, 17].

Nos travaux sont aussi liés à la décomposition intrinsèque d'images, soit la décomposition en l'ombrage diffus (ou *shading*) et la réflectance (qui donnent la partie diffuse une fois multipliées) et l'ombrage spéculaire. Beaucoup de ces méthodes considèrent la scène comme purement diffuse et estiment ainsi seulement l'ombrage diffus et la réflectance [5, 19, 24]. Cela peut mener à des défaillance des méthodes comme le montre [33]. Innamorati et al. [12] ont entraîné un réseau convolutif sur données synthétiques pour la décomposition intrinsèque, y compris la partie spéculaire. [33] améliore la méthode en ajoutant des connexions résiduelles entre le générateur et l'encoder. Cependant, leur méthode nécessite un masque de segmentation (§4).

3 Méthode de suppresion de spécularités

Le modèle dichromatique [30] stipule qu'après la réflexion sur la surface d'un objet, la lumière se divise en deux parties, à un ratio qui dépend de l'indice de réfraction du matériau. Une première partie, appelée réflexion spéculaire, est réfléchie à la surface de la même manière qu'une réflexion miroire. La deuxième, appelée réflexion diffuse, pénètre la surface et se disperse dans le corps de l'objet avant de refaire surface et d'être réfléchie. Ces deux parties de la lumière s'ajoutent alors et forment l'image après intégration de toute la lumière venant de l'hémisphère supérieure de l'objet. L'intégration étant une opération linéaire, on peut décrire l'image I par :

$$I = S + D, \tag{1}$$

où S est l'image spéculaire et D est l'image diffuse. Le problème de séparation des composantes spéculaire et diffuse consiste à estimer la partie spéculaire S et/ou la partie diffuse D, puisqu'une estimation consistente de l'un suffit à retrouver l'autre par soustraction. Nous considérons donc le problème d'estimation de D.

3.1 Principe

Pour résoudre le problème mal posé qu'est la séparation, nous proposons un générateur $\mathcal{G}_{\theta_g}$ paramétrisé par ses poids θ_g. $\mathcal{G}_{\theta_g}$ est un réseau de neurones convolutif (CNN), qui prend I en entrée et génère D. Nous entraînons $\mathcal{G}_{\theta_g}$ sur une base de données synthétiques $\mathbb{T} = \{I_i, D_i\}_{i=1}^N$ constitué de N images et les vérités terrains des images diffuses correspondantes. La création de cette base de données est discutée dans §3.2. Formellement, notre méthode se résume au problème d'optimisation suivant :

$$\hat{\theta}_g = \arg\min_{\theta_g} \frac{1}{N} \sum_{i=1}^{N} \ell\left(\mathcal{G}_{\theta_g}(I_i), D_i\right), \tag{2}$$

où ℓ est notre fonction de coût spécifiquement conçue pour le problème de séparation. ℓ s'exprime par la somme pondérée de deux termes de perte. Un de ces termes est une fonction de coût de perceptuelle ("content loss") afin de superviser l'entraînement et l'autre est une fonction de coût adverse qui aide à améliorer la précision de la suppression de spécularités. Le réseau discriminateur sur lequel est construit la perte adverse a été pensé spécifiquement pour le problème, comme le montre les résultats. Il est entraîné à détecter la bonne ou mauvaise suppression des spécularités d'une image, tandis que le réseau de suppression est entraîné à tromper le discriminateur. Ce dernier et la fonction de coût ℓ sont décrits dans §3.4 et §3.5 respectivement. Figure 2 offre une vue d'ensemble de notre méthode.

3.2 Base de données synthétiques

L'obtention des vérités terrain des composantes de réflexion étant pratiquement impossible à réaliser pour un grand nombre d'images réelles, nous entraînons notre réseau sur une base de données synthétiques. Nous avons pris soin de générer des données diverses, contenant des cas difficiles en pratique, afin de permettre à notre réseau de correctement généraliser le concept de spécularité. Nos données d'entraînement $\mathbb{T} = \{I_i, D_i\}_{i=1}^N$ sont constituées

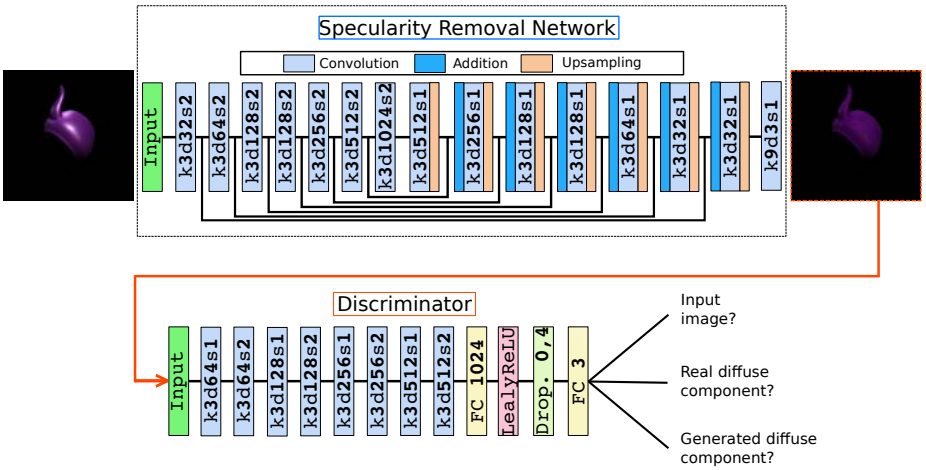

FIGURE 2 – Vue d'ensemble de notre architecture. Le réseau de suppression de spécularités prend une seule image en entrée et donne sa composante diffuse en sortie. Le réseau discriminateur intervient seulement lors de l'entraînement. Il prend une image en entrée et est entraîné à la classer parmi trois catégories : l'image d'entrée (I), la composante diffuse réelle (D) et la composante diffuse générée ($\hat{D}$) ; le générateur est entraîné à tromper le discriminateur pour que ce dernier classifie $\hat{D}$ comme D.

de $N = 20000$ images synthétisées de manière réaliste I_i et leurs images diffuses correspondantes D_i. La figure 3 montre des exemples de notre base de données.

La génération des images est automatisée par un script Blender et est réalisée avec le moteur de rendu Cycles. Chacune des N unités d'entraînement contient un seul objet, tiré parmi 8 modèle 3D, synthétisé au centre de l'image dans une orientation aléatoire. Nous avons exclu 5 modèle 3D de la base d'entraînement afin de créer également une base de données de test de 1000 images, utilisées pour l'évaluation quantitative. Nous avons paramétré la génération des images synthétiques de sorte qu'elles correspondent à une combinaison linéaire des composantes spéculaire et diffuse. La partie diffuse est modélisée par le modèle Lambertien [16] et la partie spéculaire par le modèle de Beckmann de distribution des microfacettes (Glossy BSDF dans Blender). La rugosité spéculaire est choisie aléatoirement dans la plage [0.2, 0.5]. La borne inférieure a été choisie afin de ne pas avoir de matériau s'approchant d'un miroir, ce qui dépasserait l'étendu de notre méthode et nécessiterait peut être une tout autre stratégie. Pour de tels matériaux, la scène entourant l'objet serait par exemple primordiale. La borne supérieure de 0.5 sert quant à elle à assurer la présence de spécularités en ne créant pas de matériau trop diffus. Les N unités d'entraînement de notre base sont réparties dans 4 catégories, où chacune d'entre elles sert à surmonter une difficulté.

Objets texturés aléatoirement. Ce premier ensemble contient 10000 paires d'images, synthétisées avec quatre lampes *area* dirigées vers l'objet. La position des lumières est fixée dans la scène de base et nous y ajoutons un petit déplacement aléatoire pour chaque nouveau rendu, ainsi qu'une intensité aléatoire, pour plus de diversité. Pour la même raison, nous ajoutons une texture aléatoire à chaque objet. Cela simule des objets du monde réel qui ne sont pas toujours de couleur uniforme et aide aussi le réseau à générer des images de meilleure qualité en lui permettant de travailler sur des détails fins. La base de test est également rendue avec des textures. La couleur des lumières est choisie presque blanche (deux légèrement bleues et deux autres légèrement jaunes pour imiter des lampes réelles).

Objets blancs. Si on inclut seulement la première catégorie de données, le réseau aura du mal à séparer les composantes de réflexions d'un objet blanc. En effet, le réseau apprendrait seulement à repérer et supprimer les pixels blancs de l'image vu que cela suffirait pour l'ensemble de données. On ajoute donc ce second ensemble de 4000 paires d'images montrant des objets blancs, autrement synthétisés de la même manière que la première catégorie. En forçant le réseau à reconnaître le blanc venant du matériau et le blanc venant des réflexion spéculaires, l'apprentissage est plus précis.

Lumières colorées. Ce troisième ensemble contient 2000 paires d'images, synthétisées comme le premier ensemble à la seule différence qu'on assigne une couleur aléatoire aux lampes. Cet ensemble permet de prendre en compte les cas réels où la lumière n'est pas blanche, cas qui deviennent de plus en plus fréquents avec les LEDs bien que ce soit encore peu commun (c'est pourquoi nous avons inclus seulement 2000 images). Encore une fois, cela aide le réseau à ne pas associer automatiquement les spécularités à des pixels blancs.

Environment maps. Enfin, nous avons ajouté ce dernier ensemble après observation que dans le monde réel, les spécularités sont parfois très étalées sur l'objet et non loca-

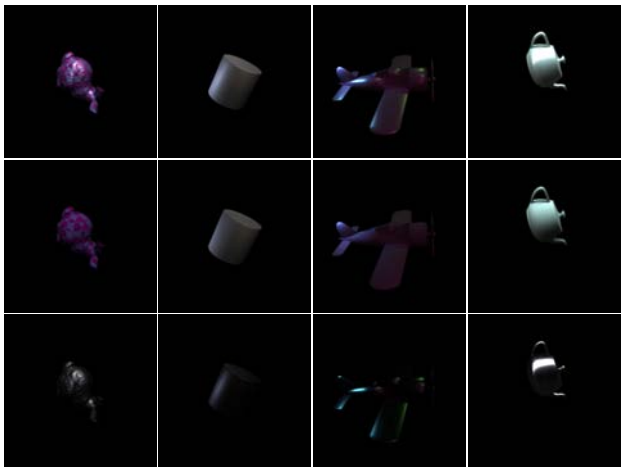

FIGURE 3 – Quatre exemples de la base de données proposée. Chaque colonne représente un ensemble parmi ceux présentés dans § 3.2 (dans le même ordre).

lisées. Cela est dû aux sources de lumières étendues et aux interréflexions, tandis que nous avons ici un objet unique dans la scène. Pour simuler ces effets, nous synthétisons cet ensemble avec des environment maps choisies parmi 6 panoramas.

3.3 Réseau de suppression de spécularités

Le réseau de suppression de spécularités est un CNN, inspiré par l'architecture du U-Net [26]. Il prend en entrée une seule image RGB, réduit sa résolution avec des convolutions de pas égal à 2, avant de la traiter puis de réaugmenter sa résolution par upsampling pour enfin générer la composante diffuse. La figure 2 (haut) illustre l'architecture de notre réseau générateur, inspirée de l'architecture de SpecularNet [23]. Plus précisément, le réseau est composé d'un bloc encodeur de sept couches convolutives avec un noyau de taille 3×3 et d'un bloc décodeur fait de huit couches de déconvolution (toutes suivies par une couche d'upsampling sauf la dernière). De plus, l'encodeur et le décodeur sont liés par des connexions résiduelles à chaque couche de convolution/déconvolution de la même taille, ce qui permet de conserver la structure spatiale de l'image d'entrée et de générer les images diffuses en haute résolution. Une activation ReLU est implémentée après chaque couche de convolution. Le réseau est entraîné avec des images de taille 256×256, mais peut ensuite être appliqué sur des images de toutes tailles puisqu'il est composé uniquement de couche de convolution.

3.4 Réseau discriminateur

Dans un GAN classique, le rôle du réseau discriminateur est de déterminer si une image est réelle ou si elle a été générée. Il s'agit donc d'un classifieur binaire entraîné à reconnaître une image réelle d'une fausse. Quant au générateur, il est entraîné à tromper le discriminateur, ce qui le pousse à générer des images de haute qualité perceptuelle, proches de la distribution des données. Dans notre cas, cette stratégie n'est pas adaptée pour deux raisons : (1) nous ne voulons pas que notre réseau reproduise la distribution de nos données (*e.g.* une image synthétique avec un objet unique au centre), qui ne rend pas compte de la diversité des images réelles ; (2) la qualité visuelle n'est pas notre souci principal, ce dernier étant de supprimer les spécularités de l'image de manière cohérente. Nous proposons donc une nouvelle stratégie d'entraînement, qui suit le paradigme du GAN de Goodfellow et al. [8], mais diffère dans le rôle du réseau discriminateur.

Optimisation adverse multi-classes. Le rôle de notre discriminateur est de différencier les images diffuses générées des images diffuses réelles, mais aussi des images d'entrée I, qui contiennent des spécularités. Il ne retourne donc pas une seule valeur en sortie, mais un tenseur de trois valeurs représentant les probabilités pour l'image d'appartenir à chacune des classes et dont la somme vaut 1. Pour cela, nous avons remplacé la fonction d'activation sigmoïde habituelle de la dernière couche par une activation softmax. Nous appelons $\mathcal{D}_{\theta_d}$ notre discriminateur et $\hat{D} = \mathcal{G}_{\theta_g}(I)$ une image diffuse générée par $\mathcal{G}_{\theta_g}$. Le réseau discriminateur est optimisé alternativement avec $\mathcal{G}_{\theta_g}$ afin de résoudre le problème adverse min-max :

$$\min_{\theta_g} \max_{\theta_d} \sum_{i=1}^{3} \mathbb{E}_{x_i \sim p_{x_i}} \left\{ \log \mathcal{D}^{(i)}_{\theta_d}(x_i) \right\} + \sum_{i=1, j=1, i \neq j}^{3} \mathbb{E}_{x_j \sim p_{x_j}} \left\{ \log \left[1 - \mathcal{D}^{(i)}_{\theta_d}(x_j) \right] \right\}, \tag{3}$$

où :

- $\mathcal{D}^{(i)}_{\theta_d}$ est la i^{eme} sortie de $\mathcal{D}_{\theta_d}$ et représente la probabilité que l'image appartienne à la classe C_i.
- x_i est une image tirée de la distribution p_{x_i} qui correspond à la classe C_i.
- $C_i \in \{C_1, C_2, C_3\} = \{C_I, C_D, C_{\hat{D}}\}$ est une des trois classes.

Notez que Eq. (3) dépend implicitement de θ_g quand $j = 3$ et $x_j = \hat{D} = \mathcal{G}_{\theta_g}(I)$. L'idée derrière ce discriminateur multi-classes est que le générateur sera entraîné à tromper trois discriminateurs au lieu d'un, ce qui consiste à se rapprocher de D et s'éloigner de I et $\hat{D}$. Cela apporte deux avantages lors de l'entraînement : (1) le réseau sera forcé à trouver les seuls caractéristiques qui différencient I et D, *i.e.* la présence de réflexions spéculaires ; (2) la classification entre D et $\hat{D}$ assure une bonne qualité visuelle et la bonne regénération de la composition de l'image.

Architecture. Le discriminateur prend en entrée l'image de taille 256×256. Une vue d'ensemble de son architecture est donnée dans la figure 2 (bas). Le réseau est constitué de huit couches convolutives avec des noyaux de taille 3×3. La résolution est réduite toutes les deux couches avec des convolutions de pas égal à 2, allant de 256×256 à 16×16, tandis que le nombre de filtres (la profondeur) augmente à

chaque couche, allant de 64 à 512. Chaque couche convolutive est suivie d'une activation LeakyReLU et d'un 'batch normalization' excepté sur la dernière couche. Nous implémentons ensuite une couche 'Flatten' puis une couche 'Dense' pour former le vecteur de dimension 3.

3.5 Fonction de coût

Pour stabiliser l'apprentissage et en même temps entraîner un réseau efficace et précis, nous définissons notre fonction de coût comme la somme d'un terme perceptuel ('content loss') et de notre fonction perte adverse spécifique au problème de séparation :

$$\ell\left(\mathcal{G}_{\theta_g}(I), D\right) = \ell_{content} + \lambda \times \ell_{separation}, \qquad (4)$$

où λ est un paramètre de régularisation afin d'amener les deux termes à la même échelle.
Nous employons l'erreur quadratique moyenne (MSE en anglais) comme fonction perceptuelle :

$$\ell_{content} = \frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H}\left(D_{x,y} - \mathcal{G}_{\theta_g}(I_{x,y})\right)^2, \qquad (5)$$

avec W et H la largeur et la hauteur de l'image d'entrée respectivement. Nous avons exploré d'autres possibilités, telles que mesurer l'erreur sur une couche intermédiaire d'un réseau discriminateur [18, 29], mais elles n'apportent rien sur nos données relativement simples.
Le terme discriminatif de la fonction de perte est défini sur les probabilités en sortie du réseau discriminateur et met à jour les poids du générateur via le gradient de :

$$\begin{aligned}\ell_{separation}(\mathcal{G}_{\theta_g}(I), D) &= \log\left\{D^{(3)}_{\theta_d}(\mathcal{G}_{\theta_g}(I))\right\} \\ &- \log\left\{D^{(1)}_{\theta_d}(\mathcal{G}_{\theta_g}(I))\right\} - \log\left\{D^{(2)}_{\theta_d}(\mathcal{G}_{\theta_g}(I))\right\},\end{aligned} \qquad (6)$$

où, pour rappel, 3, 2 et 1 correspondent à $\hat{D}$, D et I respectivement. Comparée à la fonction de perte adverse du GAN exprimée par Goodfellow et al. [8], notre expression a deux termes supplémentaires, ce qui se traduit par une rétropropagation du gradient plus forte et plus guidée.

3.6 Détails d'entraînement

Nous avons entraîné nos réseaux à partir de zéro sur carte graphique NVIDIA GeForce GTX 1070 en utilisant la base de données décrite dans § 3.2. Les réseaux générateur et discriminateur sont entraînés de manière alternative avec des batchs de taille 16. Les images d'entrée sont normalisées pour avoir des valeurs comprises dans [0,1]. Notre modèle final a été entraîné pendant $30,000$ itérations avec un taux d'apprentissage ('learning rate') de $2 \cdot 10^{-4}$ et un taux de dégradation ('decay rate') de 0 en utilisant l'optimisateur ADAM [14]. Le générateur et le discriminateur sont mis à jour alternativement à chaque itération pour résoudre le problème min-max (3). De plus, le générateur est également mis à jour pour résoudre le problème d'optimisation (2) via le gradient de la fonction (4), avec un paramètre de régularisation $\lambda = 10^{-3}$. Nous implémentons notre modèle avec le backend Keras [6].

4 Résultats expérimentaux

Dans cette partie, nous présentons les résultats de notre méthode, sur des données synthétiques comme réelles. Nous évaluons notre réseau de suppression de spécularités quantitativement sur les données synthétiques, pour lesquelles nous avons les vérités terrain, et qualitativement sur les données réelles. Nous comparons la performance de notre méthode avec trois méthodes de séparation à l'état de l'art : Tan et al. [35] et Shen et al. [32], qui sont des méthodes basées modèle, et Shi et al. [33] qui est basée apprentissage. Le code pour les deux première méthodes est disponible sur les pages web des auteurs et nous avons téléchargé le modèle de [33] sur la page GitHub de l'auteur [1]. Nos données ne convenant pas à leur méthode, et leurs données n'étant pas disponibles, nous n'avons pas pas pu faire de comparaison des deux méthodes entraînées sur les mêmes données. Nous montrons également la contribution de l'optimisation adverse multi-classes dans §4.3 et discutons les limites de notre méthode dans §4.4.

4.1 Évaluation sur données synthétiques

Nous offrons une analyse qualitative et quantitative des résultats d'estimation de la composante diffuse sur les données synthétiques. Perceptuellement, nous évaluons les différentes méthodes sur cinq exemples de la base de test (une image par modèle 3D). Nous confirmons quantitativement cette évaluation perceptuelle sur des métriques standards, L2 (MSE) et DSSIM [10], et sur des métriques développées récemment basées sur la similarité entre deux images dans l'espace caractéristique d'un réseau de neurones. En effet, Zhang et al. [38] ont montré que ces métriques fournissent une représentation des images qui s'accorde étonnamment bien avec la perception humaine. En particulier, nous considérons la métrique correspondant à une couche du réseau Squeeze [11] (notée NET) et une version linéaire de cette métrique proposée par [38] (notée LIN-NET).
La figure 4 montre les résultats d'estimation des composantes diffuses pour les différentes méthodes de référence et la nôtre, sur 5 des 1000 images de test. Visuellement, notre méthode donne des résultats indéniablement plus proches des vérités terrain. Notez que les résultats de [33] contiennent des artefacts de reconstruction sur les contours de l'image car leur méthode nécessite un masque de segmentation de l'objet pour fonctionner.
Le tableau 1 donne la comparaison quantitative entre les différentes méthodes de la littérature, notre méthode et une méthode de référence dite 'Baseline' qui correspond à un GAN binaire classique. Les erreurs reportées correspondent à une moyenne sur les 1000 images de test. Celles-ci indiquent que notre méthode surpasse celles de l'état de l'art sur toutes les métriques, s'accordant ainsi avec l'analyse qualitative. Plus particulièrement, notre méthode donne de meilleurs résultas que la Baseline, montrant ainsi l'apport de notre GAN multi-classes.

1. https ://github.com/shi-jian/shapenet-intrinsics

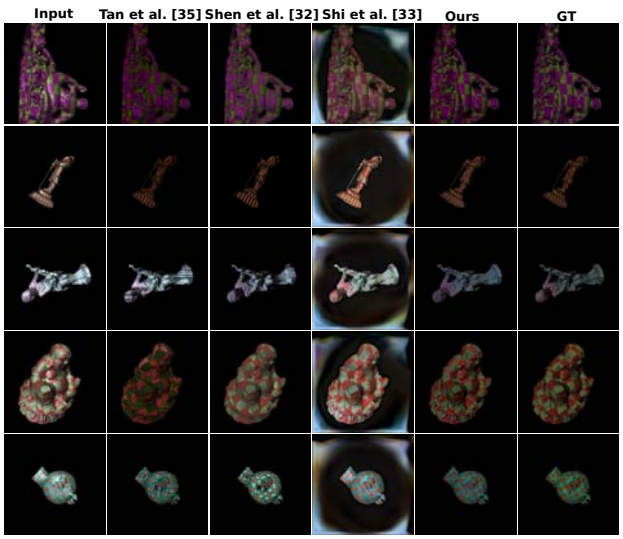

FIGURE 4 – Résultats d'estimation de la composante diffuse par les différentes méthodes. La première colonne correspond à l'image d'entrée et la dernière à la vérité terrain.

Methods	L2	DSSIM	NET	LIN-NET
Tan et al. [35]	0.020	0.052	0.408	0.054
Shen et al. [32]	0.016	0.048	0.380	0.052
Shi et al. [33]	0.222	0.410	2.363	0.313
Baseline	0.017	0.059	0.474	0.064
Ours	**0.014**	**0.035**	**0.271**	**0.034**

TABLE 1 – Résultats quantitatifs de l'estimation de la composante diffuse par les différentes méthodes. Baseline est une méthode de référence, identique à la nôtre mais entraînée avec un GAN classique (classification binaire).

Notre réseau a été entraîné sur des données synthétiques, similaires aux images test, nous pouvons donc penser qu'il est normal qu'il surpasse les autres méthodes. Toutefois, nous montrons par la suite que celui-ci est également cohérent sur des images réelles, démontrant la pertinence de notre architecture et de notre base de données.

4.2 Évaluation sur données réelles

Dans cette partie, nous évaluons notre méthode sur des images réelles. Les résultats peuvent être vus dans la figure 5. Notre réseau est capable de supprimer les spécularités d'images diverses, allant de celles qui ressemblent à nos données avec un fond noir à des scènes complexes comme les objets en bois ou le ballon Terre. Au vu de nos images d'entraînement assez simples et du fait qu'aucune géométrie vue dans les images réelles ne s'y trouvait, cela atteste de la bonne généralisation de notre réseau. La méthode [33] quant à elle ajoute des artifacts à la génération et ne parvient pas à modifier seulement les pixels spéculaires. Nous attribuons cela à notre réseau discriminateur qui contraint le générateur à capturer la distribution des spécularités et

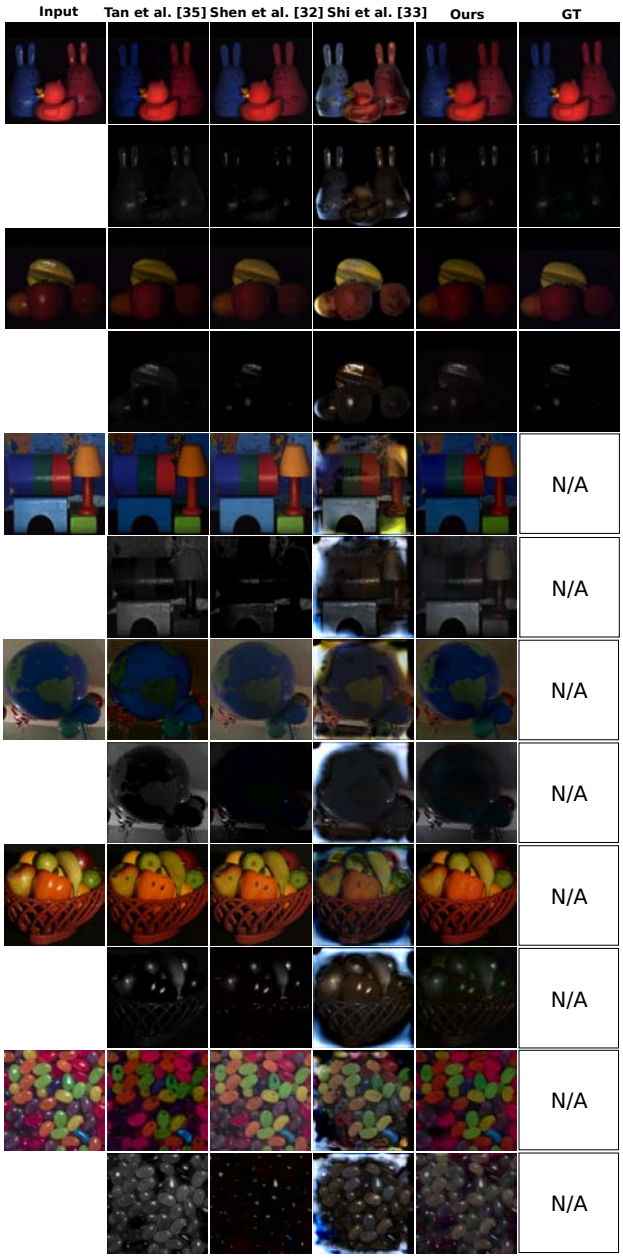

FIGURE 5 – Résultats et comparaison de notre méthode sur des images réelles. L'image d'entrée est sur la 1ère colonne et pour chaque image, la ligne du haut correspond à la composante diffuse et la ligne du bas à la composante spéculaire. Les vérités terrain sont sur la dernière colonne quand elles sont disponibles. Notre composante spéculaire est obtenue en soustrayant la partie diffuse à l'image d'entrée.

l'aide à les supprimer en spécifiant les différences entre I et D. Perceptuellement, notre méthode surpasse systématiquement [33].
En comparaison visuelle avec les méthodes basées modèle, nous sommes un peu moins précis que Shen et al. [32] sur le premier et le deuxième exemples (animaux et fruits) et un peu moins précis que Tan et al. [35] sur le premier

et troisième exemples (animaux et objets en bois). Cela peut s'expliquer par le fait que ce sont des images prises en conditions de laboratoire, *i.e.* dans un environnement contrôlé, et qu'elles respectent quasi parfaitement les hypothèses émises par leur modèle : une illumination unique, parfaitement blanche, pas de pixels saturés et pas de non linéarité introduite par la caméra. Lorsque ces conditions sont rencontrées, le modèle dichromatique [30] est vérifié expérimentalement, ce n'est donc pas étonnant que des méthodes construites dessus performent bien. Cependant, bien que subjectif car nous n'avons pas de métrique adaptée pour la suppresion de spécularités, notre méthode est meilleure sur les autres exemples. Sur la cinquième image (panier de fruits), nous voyons clairement que les méthodes basées modèle ne parviennent pas à récupérer la partie diffuse à cause des pixels saturés qui ne montrent pas de couleur diffuse sous-jacente. Cela résulte en des 'trous' dans l'image où les pixels ont été totalement effacés, tandis que notre méthode les remplit de manière globalement cohérente. Il en va de même pour le dernier exemple (les bonbons). En résumé, notre méthode surpasse les autres sur ces exemples.

4.3 Contribution de l'optimisation adverse multi-classes

La figure 6 fournit les courbes d'apprentissage de notre méthode ainsi que de la Baseline (GAN classique), entraînées toutes deux avec les mêmes paramètres. Les amplitudes d'oscillation des deux courbes nous indiquent clairement que la fonction de perte multi-classes permet un apprentissage plus stable, l'instabilité étant une caractéristique connue des structures adverses. Les courbes montrent ainsi une convergence plus rapide de notre architecture et plus précise, ce qui se voit dans les images générées par chacune des méthodes (bas de la figure 6).

4.4 Limites

Notre modèle n'est évidemment pas parfait et peut échouer en application sur des images réelles. Premièrement, comme mentionné plus haut, il ne gère pas les surfaces miroir, qui nécessitent une tout autre définition du problème (données et a prioris différents). Le seul moyen de reconnaître une telle surface est d'avoir le contexte de la scène entourant l'objet. De plus, malgré nos efforts sur la base de données et le processus d'entraînement, le réseau tend à asssombrir les images entièrement, ce qui signifie qu'il modifie des pixels qui ne sont pas spéculaires, bien que les changements soient moindres par rapport aux vrais pixels spéculaires. Cet effet est visible sur les exemples objets en bois et ballon Terre de la figure 5. Cela peut s'expliquer par la simplicité de nos données, qui ne contiennent pas de contexte qui aurait pu apporter des informations précieuses sur l'illumination de la scène. Le réseau n'est donc pas entraîné pour cette tache et est forcé de se concentrer sur des indices visuels simples tels que la couleur des spécularités. Nous avons essayé d'ajouter un fond blanc afin d'aider

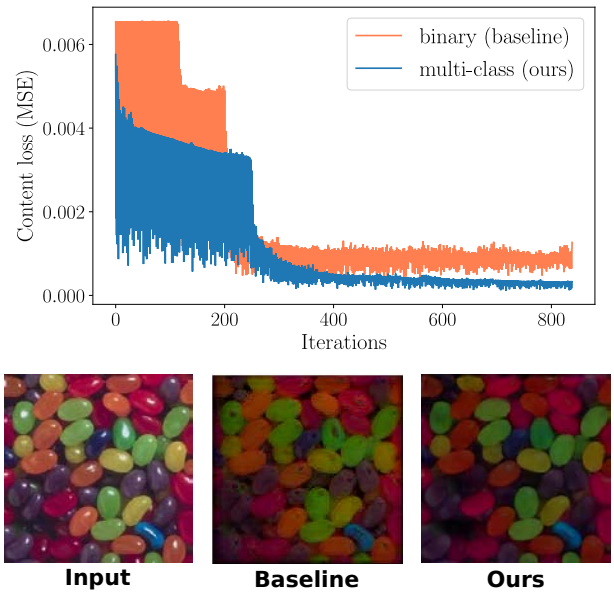

FIGURE 6 – (Haut) Courbes d'apprentissage d'un GAN classique et de notre GAN multi-classes. (Bas) Composantes diffuses générées par les deux méthodes.

le réseau à faire la différence entre des matériaux blancs et des spécularités blanches, sans observer d'amélioration nette. Nous avons également essayé l'implémentation du patchGAN qui n'améliore pas les résultats.

5 Conclusion

Nous avons proposé une nouvelle méthode de suppression de spécularités, basée apprentissage, qui capture mieux les relations complexes entre l'objet, l'illumination et l'image que les approches existantes. Notre approche tire profit de la génération de données synthétiques comme solution pour facilement avoir un grand nombre d'images. Nous avons généré ces données pour qu'elles incluent des cas difficiles de séparation, afin d'affiner la précision du générateur et l'aider à mieux généraliser le concept de spécularité. Avec le même objectif, nous avons entraîné notre modèle dans la structure GAN mais que nous avons modifié et adapté au problème de séparation des composantes de réflexion. Plus précisément, la fonction de perte adverse est définie sur un discriminateur multi-classes et non binaire, ajoutant ainsi deux termes au gradient de rétropropagation. Cela aboutit à un réseau capable de supprimer les spécularités à partir d'une seule image, sans aucune connaissance sur la scène. Nous avons évalué notre modèle sur des données synthétiques et réelles, démontrant ainsi une cohérence plus constante sur des scènes diverses par rapport à l'état de l'art.

Pour de futurs travaux, il serait intéressant d'étudier la cohérence temporelle de notre méthode pour une application éventuelle en temps réel sur des vidéos. Évidemment, notre modèle profiterait grandement de données réelles.

Références

[1] E. H. Adelson and A. P. Pentland. The perception of shading and reflectance. *Perception as Bayesian inference*, pages 409–423, 1996.

[2] D. An, J. Suo, X. Ji, H. Wang, and Q. Dai. Fast and high quality highlight removal from a single image. *arXiv preprint arXiv :1512.00237*, 2015.

[3] R. Bajcsy, S. W. Lee, and A. Leonardis. Detection of diffuse and specular interface reflections and inter-reflections by color image segmentation. *International Journal of Computer Vision*, 17(3) :241–272, 1996.

[4] J. T. Barron and J. Malik. Shape, illumination, and reflectance from shading. *IEEE transactions on pattern analysis and machine intelligence*, 37(8) :1670–1687, 2015.

[5] A. S. Baslamisli, H.-A. Le, and T. Gevers. Cnn based learning using reflection and retinex models for intrinsic image decomposition. *ArXiv e-prints*, 2017.

[6] F. Chollet et al. Keras, 2015.

[7] I. Funke, S. Bodenstedt, C. Riediger, J. Weitz, and S. Speidel. Generative adversarial networks for specular highlight removal in endoscopic images. In *Medical Imaging 2018 : Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 10576, page 1057604. International Society for Optics and Photonics, 2018.

[8] I. Goodfellow. Nips 2016 tutorial : Generative adversarial networks. *arXiv preprint arXiv :1701.00160*, 2016.

[9] K. Hara, K. Nishino, and K. Ikeuchi. Determining reflectance and light position from a single image without distant illumination assumption. In *null*, page 560. IEEE, 2003.

[10] A. Hore and D. Ziou. Image quality metrics : Psnr vs. ssim. In *Pattern recognition (icpr), 2010 20th international conference on*, pages 2366–2369. IEEE, 2010.

[11] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. Squeezenet : Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. *arXiv preprint arXiv :1602.07360*, 2016.

[12] C. Innamorati, T. Ritschel, T. Weyrich, and N. J. Mitra. Decomposing single images for layered photo retouching. In *Computer Graphics Forum*, volume 36, pages 15–25. Wiley Online Library, 2017.

[13] J. Jachnik, R. A. Newcombe, and A. J. Davison. Real-time surface light-field capture for augmentation of planar specular surfaces. In *Mixed and Augmented Reality (ISMAR), 2012 IEEE International Symposium on*, pages 91–97. IEEE, 2012.

[14] D. P. Kingma and J. Ba. Adam : A method for stochastic optimization. *arXiv preprint arXiv :1412.6980*, 2014.

[15] G. J. Klinker, S. A. Shafer, and T. Kanade. The measurement of highlights in color images. *International Journal of Computer Vision*, 2(1) :7–32, 1988.

[16] J. H. Lambert. *Photometria sive de mensura et gradibus luminis, colorum et umbrae*. Klett, 1760.

[17] E. H. Land and J. J. McCann. Lightness and retinex theory. *Josa*, 61(1) :1–11, 1971.

[18] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint*, 2016.

[19] L. Lettry, K. Vanhoey, and L. van Gool. Deep unsupervised intrinsic image decomposition by siamese training. *arXiv preprint arXiv :1803.00805*, 2018.

[20] S. Lin and S. W. Lee. Estimation of diffuse and specular appearance. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 855–860. IEEE, 1999.

[21] S. Lin, Y. Li, S. B. Kang, X. Tong, and H.-Y. Shum. Diffuse-specular separation and depth recovery from image sequences. In *European conference on computer vision*, pages 210–224. Springer, 2002.

[22] S. Lin and H.-Y. Shum. Separation of diffuse and specular reflection in color images. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2001.

[23] A. Meka, M. Maximov, M. Zollhoefer, A. Chatterjee, C. Richardt, and C. Theobalt. Live intrinsic material estimation. *arXiv preprint arXiv :1801.01075*, 2018.

[24] T. Narihira, M. Maire, and S. X. Yu. Direct intrinsics : Learning albedo-shading decomposition by convolutional regression. In *Proceedings of the IEEE international conference on computer vision*, pages 2992–2992, 2015.

[25] S. K. Nayar, X.-S. Fang, and T. Boult. Separation of reflection components using color and polarization. *International Journal of Computer Vision*, 21(3) :163–186, 1997.

[26] O. Ronneberger, P. Fischer, and T. Brox. U-net : Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[27] C. Rother, M. Kiefel, L. Zhang, B. Schölkopf, and P. V. Gehler. Recovering intrinsic images with a global sparsity prior on reflectance. In *Advances in neural information processing systems*, pages 765–773, 2011.

[28] Y. Sato and K. Ikeuchi. Temporal-color space analysis of reflection. *JOSA A*, 11(11) :2990–3002, 1994.

[29] M. E. A. Seddik, M. Tamaazousti, and J. Lin. Generative collaborative networks for single image super-resolution. *arXiv :1902.10467*, 2019.

[30] S. A. Shafer. Using color to separate reflection components. *COLOR Research & Application*, 10(4) :210–218, 1985.

[31] H.-L. Shen and Q.-Y. Cai. Simple and efficient method for specularity removal in an image. *Applied optics*, 48(14) :2711–2719, 2009.

[32] H.-L. Shen, H.-G. Zhang, S.-J. Shao, and J. H. Xin. Chromaticity-based separation of reflection components in a single image. *Pattern Recognition*, 41(8) :2461–2469, 2008.

[33] J. Shi, Y. Dong, H. Su, and X. Y. Stella. Learning non-lambertian object intrinsics across shapenet categories. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 5844–5853. IEEE, 2017.

[34] R. T. Tan and K. Ikeuchi. Estimating chromaticity of multi-colored illuminations. In *Workshop on Color and Photometric Methods in Computer Vision*, volume 2. Citeseer, 2003.

[35] R. T. Tan and K. Ikeuchi. Separating reflection components of textured surfaces using a single image. *IEEE transactions on pattern analysis and machine intelligence*, 27(2) :178–193, 2005.

[36] Y. Weiss. Deriving intrinsic images from image sequences. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, pages 68–75. IEEE, 2001.

[37] L. B. Wolff and T. E. Boult. Constraining object features using a polarization reflectance model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(7) :635–657, 1991.

[38] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. *arXiv preprint*, 2018.

[39] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*, 2017.

[40] M. Zwicker. Specular-to-diffuse translation for multi-view reconstruction.