Le transfert de fond de teint n'est pas qu'une copie de couleur

Neil Farmer, Catherine Soladié, Gabriel Cazorla, Renaud Séguier

Le transfert de fond de teint n'est pas qu'une copie de couleur

Neil Farmer[1,2] Catherine Soladié[2] Gabriel Cazorla[1] Renaud Séguier[2]

[1] Chanel Parfums Beauté, Innovation Research and Development, Pantin, France
[2] CentraleSupélec, IETR UMR CNRS 6164, France

neil.farmer@chanel.com

Résumé

Le transfert de maquillage vise à appliquer un maquillage donné (extrait d'un visage appelé référence) sur un visage non maquillé (source) tout en préservant les attributs d'identité de la source. Les méthodes récentes tendent à considérer le transfert de maquillage comme une copie de la couleur. Cependant, cette hypothèse conduit à appliquer un maquillage esthétique sur une image réaliste au lieu de générer un maquillage réaliste. Par conséquent, même si le maquillage généré peut sembler réaliste, il n'est pas conforme à la réalité. Notre approche vise à préserver les informations relatives au teint du visage source. De plus, pour éviter que notre modèle ne mélange les processus de maquillage et de démaquillage, ces processus sont séparés. Les résultats quantitatifs montrent que nous sommes plus performants que les architectures de transfert de maquillage de l'état de l'art, tant pour la précision du fond de teint que pour sa cohérence. Ces résultats sont confirmés par une étude utilisateurs. Le code et le modèle entrainé sont disponibles sur GitHub.

Mots Clef

GAN, Transfert de Maquillage, Préservation d'Identité.

Abstract

Makeup Transfer aims to apply a given makeup style (extracted from an image called reference) onto a non-makeup face (source) while preserving identity attributes. Makeup transfer is treated as two distinct subtasks : transferring makeup from the reference face to the source face, and removing makeup from the reference face. Recent methods have considered makeup transfer as a color copying task. However, this assumption led to the generation of good-looking makeup on realistic images instead of realistic makeup. As a result, even though the generated makeup may appear realistic, it is far from the true render of the reference makeup on the skin of the source face. In this paper, the proposed approach aims to preserve the source face's skin tone information. In addition, to prevent our model from entangling the makeup and makeup removal processes, the model separates the makeup removal and makeup addition tasks. The quantitative results show that we outperform state-of-the-art makeup transfer architectures for both foundation accuracy and consistency. These results are further confirmed by a crowd-sourced user study.

Keywords

GAN, Makeup Transfer, Identity Preservation.

1 Introduction

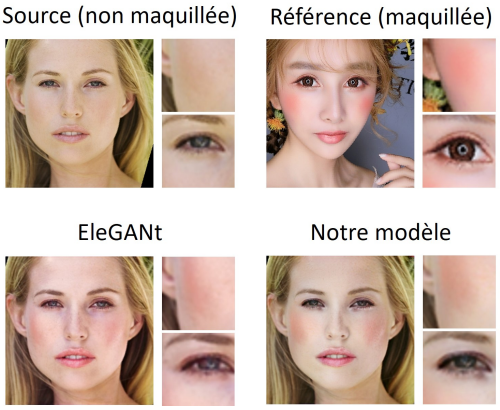

FIGURE 1 – Exemple de transfert de maquillage avec EleGANt [1] et notre méthode. La source est l'image à maquiller. La référence est l'image avec le maquillage cible. Trois zones de maquillage peuvent être définies : les lèvres, le contour des yeux et le visage.

Le transfert de maquillage vise à extraire le maquillage d'un visage référence et à l'appliquer sur une source non maquillée [2]. Ces dernières années, de nombreuses applications ont été développées pour améliorer l'expérience client. La prévisualisation du maquillage, ou le démaquillage virtuel sont des exemples de ces applications [3]. Les implémentations de ces applications se font sur des dispositifs tels que des miroirs connectés ou des smartphones [4]. Notre recherche porte sur le transfert réaliste du teint de la peau lors de l'application du maquillage. Les axes principaux de recherche dans le domaine du transfert de maquillage portent sur l'alignement des parties du visage entre la source et la référence [5, 6, 7, 8], et sur le maquillage modifiable localement [1]. Cependant, peu d'at-

tention a été accordée au réalisme du maquillage. Alors que les méthodes récentes considèrent le transfert du fond de teint comme une copie de couleur, nous considérons plutôt la couleur du fond de teint comme une combinaison de la couleur de la peau et de la couleur du maquillage. Cette hypothèse repose sur le fait que le fond de teint n'est pas un produit totalement opaque [9] et que le teint de la peau affecte donc la couleur perçue. Par conséquent, la copie des couleurs du fond de teint de la référence vers la source donne un fond de teint irréaliste dans l'image générée (Figure 1).
Pour répondre à ces problématiques, nous proposons les contributions suivantes :

- Une nouvelle pseudo vérité terrain (ou Pseudo Ground Truth ou PGT), c'est-à-dire une image synthétisée indépendamment du générateur et utilisée comme cible d'entraînement, basée sur l'α blending pour préserver l'identité et, plus précisément, la couleur de la peau de la source dans le PGT,
- Une architecture à double décodeur, l'un pour le maquillage et l'autre pour le démaquillage,
- Une nouvelle fonction de perte (*loss function*) qui garantit que les zones non maquillées (les cheveux, l'arrière-plan ou la couleur des yeux) restent inchangées.

Grâce aux données collectées dans des études antérieures, nous avons développé une base de données privée contenant des visages avec et sans fond de teint, ainsi que la référence du fond de teint porté. Cette base de données nous a permis de mesurer le réalisme et la cohérence du fond de teint transféré, c'est-à-dire l'écart entre toutes les images générées pour une même source et différentes références portant un même fond de teint.

2 Etat de l'art

2.1 GANs pour le Transfert de Style

Les GANs [10] ont été largement utilisés pour une variété d'applications : génération d'images [11, 12], transfert de style [13], manipulation d'images [14, 15], et transfert de maquillage [5, 6, 1, 16, 17, 18, 19, 20, 21]. L'idée principale qui a fait le succès des GANs est l'apprentissage antagoniste, qui oblige le générateur à produire des images indiscernables de l'ensemble d'apprentissage. Pour la tâche de transfert de maquillage, l'architecture CycleGAN [22] est largement utilisée dans la communauté pour sa capacité à transférer les images d'un domaine X à un domaine Y à partir de paires non labélisées d'images du domaine X et Y. Ceci est réalisé en utilisant un GAN pour apprendre le transfert de X vers Y et un second GAN pour apprendre le processus inverse (transfert de Y vers X). En utilisant ces deux générateurs, les auteurs ont introduit une perte de cohérence du cycle qui oblige les générateurs à produire une image qui ne peut être distinguée de la distribution cible apprise et qui préserve la sémantique de l'image d'entrée.

2.2 Transfert de Maquillage

Les tâches de transfert de maquillage sont étudiées depuis plus d'une décennie à l'aide de techniques traditionnelles de traitement d'images [23, 24, 25] ou de techniques d'apprentissage profond [5, 6, 1, 16, 17, 18, 19, 20, 26, 27, 28, 29]. Suite au succès de CycleGAN [22] pour le transfert de domaine d'image à image, cette architecture a été rapidement adoptée pour le transfert de maquillage [5, 6, 1, 16, 28, 30]. Cependant, Tingting Li et al. [16] ont noté que le transfert de maquillage est un transfert de style au niveau de l'instance, alors que CycleGAN est un transfert de style au niveau global. En d'autres termes, CycleGAN transfère une représentation globale du maquillage au lieu d'un maquillage spécifique. Pour résoudre ce problème, ils proposent BeautyGAN [16] qui introduit une nouvelle fonction de perte qui ajoute une supervision supplémentaire pour transférer un maquillage spécifique. Yang et al. ont amélioré l'idée avec EleGANt [1] en ajoutant une pseudo vérité terrain (noté PGT pour Pseudo Ground Truth) qui guide le générateur pour préserver la nuance du fond de teint et les détails spatiaux du maquillage, tels que le blush. Les architectures précédentes ([5, 6, 1, 16, 31]) utilisent la correspondance d'histogramme pour transférer la couleur de fond de teint de la référence vers la source. Le principal problème de la méthode de correspondance des histogrammes est qu'elle transfère la couleur de la référence sans tenir compte de la couleur de peau de la source, ce qui entraîne le transfert non seulement de la couleur du fond de teint, mais aussi de la couleur de peau du visage de référence, des ombres, etc.
Un autre aspect important des recherches antérieures est le démêlage des caractéristiques [32, 33, 34]. BeautyGlow [17] propose un modèle dérivé de Glow pour démêler les caractéristiques en deux vecteurs latents distincts (l'un pour le maquillage et l'autre pour les caractéristiques d'identité). Pour ce faire, Chen et al.[17] utilisent une fonction inversible constituée d'une séquence de matrices de transformation. Une supervision est ajoutée pour aider le réseau à distinguer les caractéristiques du maquillage de celles d'identité. Pour ce faire, on suppose que l'espace des caractéristiques contient à la fois des caractéristiques d'identité et de maquillage. Ainsi, si l'on supprime les caractéristiques d'identité de l'espace des caractéristiques, il ne reste que les caractéristiques du maquillage. SSAT [5] propose d'utiliser les informations sémantiques pour rendre le style de maquillage à la position sémantiquement correspondante de l'image de référence. Dans ce travail, nous améliorons l'idée de PGT en remplaçant la correspondance d'histogramme [35] par un α blending pour préserver le teint original de la source dans le PGT (3.3). De plus, nous ajoutons une nouvelle fonction de perte pour réduire les changements dans l'arrière-plan des zones non maquillées (3.5). Nous partons du principe que les processus de maquillage et de démaquillage sont deux tâches différentes et nécessitent donc deux processus distincts (3.4).

3 Méthode

Après avoir présenté les formulations utilisées (3.1) et l'architecture globale de notre modèle (3.2), nous décrivons chacune des principales contributions : la modification de la pseudo vérité terrain (PGT) pour la préservation de l'identité (3.3), le double décodeur pour le démêlage des processus (3.4) et la nouvelle fonction de perte pour la cohérence de l'image (3.5).

3.1 Formulation

Soit $X \subset \mathbb{R}^{H \times W \times 3}$ l'ensemble des images non maquillées et $Y \subset \mathbb{R}^{H \times W \times 3}$ l'ensemble des images maquillées, où H et W sont les constantes définies respectivement pour la hauteur et la largeur de l'image. Étant donné une image source $x \in X$ et une image de référence $y \in Y$, l'objectif du processus de maquillage est d'apprendre une fonction de transfert G telle que : $\hat{x} = G(x, y)$ où $\hat{x}$ a le style de maquillage de la référence y tout en préservant l'identité de la source x.
A l'inverse, l'objectif du processus de démaquillage est d'apprendre la fonction de transfert F tel que : $\hat{y} = F(y, x)$ où $\hat{y}$ a le style de maquillage de la source x, donc sans maquillage, tout en préservant l'identité de la référence y. De plus, nous définissons la source reconstruite x_{rec} comme suit : $x_{rec} = F(\hat{x}, \hat{y})$ et la référence reconstruite y_{rec} comme : $y_{rec} = G(\hat{y}, \hat{x})$.

3.2 Architecture du Modèle

L'architecture globale du réseau est présentée dans la Figure 2. Elle consiste en deux étapes d'apprentissage distinctes : l'apprentissage de l'encodeur/décodeur et l'apprentissage auto-supervisé du générateur α. L'apprentissage de l'encodeur/décodeur consiste à :

1. Extraire les caractéristiques avec un encodeur E, prenant en entrée une source x et une référence y.
2. Générer des images avec un décodeur pour le maquillage et le démaquillage (notés respectivement D_m et D_{mr}).

En plus des fonctions de perte utilisées par EleGANt [1], nous avons ajouté la fonction de perte *Makeup Free Area Loss*, qui assure la cohérence des zones sans maquillage (comme l'arrière-plan), comme décrit dans 3.5.

3.3 Pseudo Vérité Terrain (PGT)

Pour transférer le fond de teint de la référence y à la source x, les méthodes de l'état l'art [5, 6, 1, 16] considèrent le transfert comme une simple copie de couleur du teint de la référence vers la source sans prise en compte de leur teint. Nous considérons au contraire que le fond de teint n'est pas un produit entièrement opaque [9] et que le teint d'un visage maquillé est la somme de son teint et de la couleur du fond de teint. Par conséquent, il est crucial de prendre en compte la couleur de la peau dans les tâches de maquillage et de démaquillage. Pour répondre à ces nouvelles hypothèses, nous proposons de remplacer la technique de correspondance des histogrammes par la méthode d'α blending [36]. Cela permet de prendre en compte la couleur de peau originale de la source et la couleur du fond de teint de la référence (Figure 3).
La méthode traditionnelle d'α blending combine deux images avec l'opacité d'une image à $\alpha \in [0, 1]$ et l'opacité de la seconde image à $1 - \alpha$. Les images sont ensuite additionnées avec leur nouvelle opacité.
Cependant, la technique d'α blending permet d'obtenir uniquement des couleurs comprises entre les deux images. Toutefois, dans la tâche de transfert de maquillage, le fait de restreindre le blending entre les couleurs de la source et les couleurs de la référence aboutirait dans certains cas à un rendu irréaliste de la couleur du fond de teint. Par exemple, pour un fond de teint qui doit éclaircir le teint de la source, le PGT doit éclaircir le teint de la peau même si le teint de peau de la référence est plus foncé. Pour cela, on modifie l'α blending :

$$PGT_x = (1.5 \times \alpha_x + 0.5) \cdot x - (1.5 \times \alpha_x - 0.5) \cdot y \quad (1)$$

où x et y sont les images de la source et de la référence, respectivement, et $\alpha_x \in \mathbb{R}^{1 \times H \times W \times 3 \times [-1,1]}$ est un masque de mélange renvoyé par un réseau entraîné. On considère également que le maquillage affecte chaque canal de couleur et chaque pixel indépendamment des autres. Ainsi l'α blending est appliqué à chaque canal et chaque valeur de pixel avec des valeurs α indépendantes. Le masque α a ainsi les mêmes dimensions que les images source/référence. Comme mentionné ci-dessus, les α sont spécifiques à la localisation et doivent donc être appris (partie verte dans la figure 2). Cela se fait à l'aide d'un deuxième réseau. Ce réseau permet d'obtenir une carte $1 \times H \times W \times 3$ pour toute paire d'images source/référence avec une hauteur H, une largeur W, et 3 canaux. Pour guider ce réseau, nous l'entraînons à reconstruire les images originales à partir des images générées (c'est-à-dire à reconstruire la source x et la référence y à partir de $\hat{x}$ et $\hat{y}$). Pour ce faire, nous supposons qu'il existe un masque de fusion $\alpha_{rec\_x,y}$ qui conduirait à la reconstruction de la source (x) après la fusion de $\hat{x}$ et $\hat{y}$. Avec cette hypothèse, nous pouvons trouver les α optimaux qui conduisent à reconstruire les données originales en utilisant les données générées. Ces valeurs optimales α_{rec_x} sont déterminées de la façon suivante :

$$\alpha_{rec_x} = \frac{x - 0.5 \cdot \hat{x} - 0.5 \cdot \hat{y}}{1.5 \cdot (\hat{x} - \hat{y}) + \epsilon} \quad (2)$$

où x est l'image source, $\hat{x}$ et $\hat{y}$ sont les images générées et ϵ est une constante égale à $1e^{-6}$. Enfin, la différence absolue entre le masque optimale α_{rec_x} et le masque générée $\alpha_{rec_{\hat{x}}}$ est rétro-propagée à travers le générateur α.
Comme dans l'architecture principale, nous avons utilisé, pour le générateur α, une architecture encodeur/décodeur similaire à l'architecture principale. Nous considérons également que le fond de teint tend à cacher les défauts de la peau (rides, tâches de rousseur, etc.) [25]. À cette fin, nous utilisons le filtre guidé rapide (FGF) pour créer le PGT maquillé PGT_x[37]. Enfin, pour le rouge à lèvres et le contour

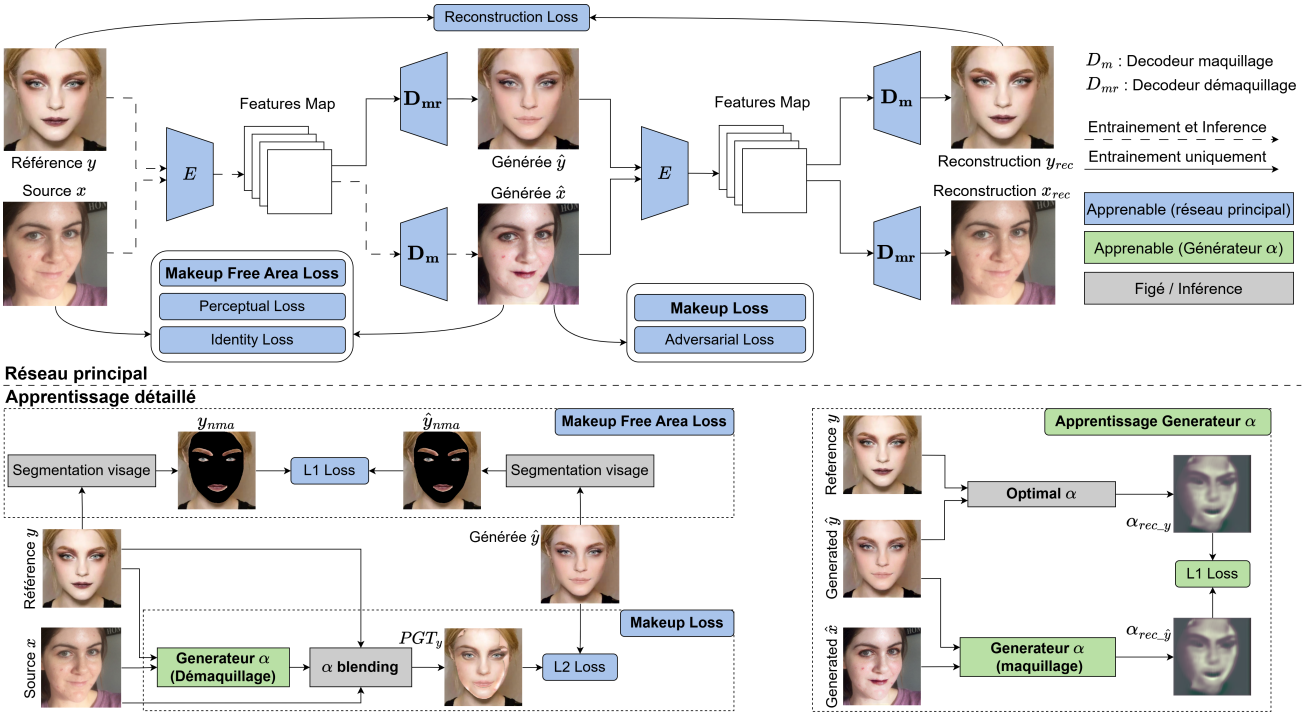

FIGURE 2 – Architecture du réseau pour le transfert de maquillage. Les principaux modules du réseau et les fonctions de perte sont représentés en bleu. Les modules et les fonctions de perte du générateur α sont représentés en vert. Toutes les fonctions de perte sont appliquées aux tâches de maquillage et de démaquillage. La partie supérieure détaille l'apprentissage global du réseau. La partie inférieure détaille les principales contributions.

des yeux, nous supposons que ces deux produits sont plus opaques que le fond de teint. Cette hypothèse conduit à conserver la correspondance des histogrammes pour ces deux produits comme dans EleGANt [1].

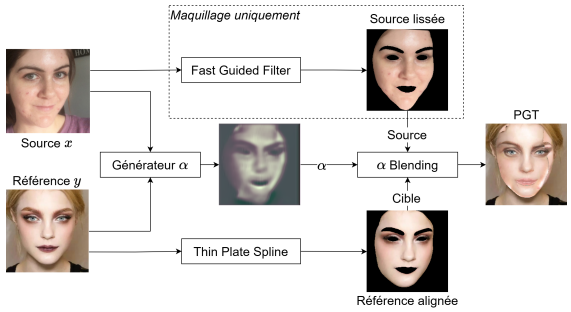

FIGURE 3 – La pseudo-vérité terrain (PGT) est le résultat d'un α blending [36] entre la référence et la source (lissée pour le processus de maquillage). Le blending est calculé à partir d'un masque déterminé pour chaque pixel du visage. Ce masque est obtenu à l'aide d'un réseau auto-supervisé appelé générateur α.

3.4 Double Décodeur

Nous considérons le maquillage et le démaquillage comme deux tâches distinctes. Alors que le maquillage ajoute des attributs de maquillage aux caractéristiques d'identité, le démaquillage est plutôt l'inverse, visant à augmenter les attributs d'identité (tâches de rousseur, rides, etc.) et d'enlever les attributs de maquillage. Cette observation nous conduit à utiliser deux décodeurs différents sans partage de poids, un pour chaque tâche (Figure 2), au lieu d'un décodeur global comme dans EleGANt [1].

3.5 Makeup-free area loss

Pour garantir la non-modification des zones non maquillées (les cheveux, l'arrière-plan et les yeux), nous introduisons une nouvelle fonction de perte appelée *Makeup Free Area Loss*. Soit $M \subset \mathbb{Z}_2^{H \times W \times S}$ un ensemble de S masques contenant une segmentation du visage, avec $\mathbb{Z}_2 = \{0, 1\}$, et $A \subset S$ l'ensemble des valeurs sémantiques des zones du visage qui peuvent être maquillées (la zone du visage excluant les yeux, les sourcils et les dents). Nous définissons la *Makeup Free Area Loss* (MFA) comme suit :

$$\mathcal{L}_{MFA} = \|x^{nma} - \hat{x}^{nma}\|_1 \tag{3}$$

où $x^{nma} = x \odot M_{a \notin A}$ et $\hat{x}^{nma} = \hat{x} \odot M_{a \notin A}$ sont les zones non maquillées (nma), c'est-à-dire les zones qui doivent rester inchangées pour, respectivement, l'image source et l'image maquillée transférée. $\|\|_1$ est la norme L1. Dans nos expériences, ces zones sont définies à l'aide de la segmentation obtenue avec un BiseNetV2 [38] pré-entraîné sur CelebAMask-HQ [39]. Les valeurs sémantiques A sont le visage et le nez.

4 Résultat

4.1 Jeu de données et apprentissage

Dans cette section, nous présentons les paramètres utilisés pour l'entrainement et les tests. Pour la première fois dans le domaine du transfert de maquillage, nous avons pu comparer les images générées avec leur vérité terrain. Nous avons évalué quantitativement nos résultats à l'aide de deux nouvelles mesures : la précision et l'erreur de cohérence du fond du teint. En outre, nous avons évalué nos résultats à l'aide de mesures traditionnelles selon trois critères : la conservation de l'identité, la qualité de l'image et la préservation des détails.

Jeu d'apprentissage Nous avons entraîné notre modèle sur le jeu de données Makeup Transfer dataset (MT) créé par Li et al. [16] et largement utilisé par la communauté [16, 1, 5, 6]. Le jeu de données contient 2719 images de maquillage et 1115 images non maquillées. Elle comprend diverses poses faciales, styles de maquillage et expressions.

Détails d'entrainement Pour la phase d'entraînement, nous suivons la répartition entrainement-test d'EleGANt : 250 images non maquillées et 100 images maquillées sont choisies aléatoirement pour le test et retirées du jeu d'entrainement. Ce jeu de test est utilisé pour les résultats qualitatif (tableau 3 et tableau 2) ainsi que pour l'étude des impactes des contributions (tableau 1). Le GAN est entraîné pendant 50 epoch avec une taille de batch de 1, sur une GPU Nvidia RTX 3080Ti Laptop 16Gb. Le taux d'apprentissage de tous les réseaux (Générateur α (3.3), Encodeur/Décodeur (3.4) et Discriminateur) est fixé à $2e^{-4}$.

Jeu de test Pour l'étude qualitative, nous avons utilisé la même séparation jeu d'entrainement/jeu de test que celui utilisé par EleGANt [1]. Actuellement, il n'existe pas de base de données publique de transfert de maquillage pour une évaluation quantitative grâce à des vérités terrain. C'est pourquoi nous avons créé une base de données de vérités terrain à partir d'images collectées antérieurement, lors de l'évaluation du fond de teint. Elle contient 139 paires de personnes sans maquillage et avec fond de teint, avec un identifiant unique par personne et par fond de teint. Il s'agit de 82 personnes avec 19 fonds de teint différents. Cet ensemble de données n'est pas accessible au public.

Comparaison quantitative Nous voulons mesurer quantitativement le réalisme des images générées. Pour cela, nous définissons deux nouvelles métriques basées sur la distance de Wasserstein [40, 41]. La première métrique est la **précision du fond de teint** : elle mesure la proximité d'une image générée par rapport à sa vérité terrain. Pour ce faire, on calcule la similarité des couleurs (à l'aide de la distance de Wasserstein) entre ces deux images. Pour éviter tout désalignement entre les images, la distance est calculée entre cinq parties du visage (joue gauche, joue droite, front, nez et menton). La deuxième mesure est l'erreur de cohérence : elle mesure l'écart entre plusieurs images générées à partir de la même source et plusieurs références portant le même maquillage. Pour ce faire, on calcule la similarité (distance de Wasserstein) entre chaque combinaison d'images générées. L'objectif de cette mesure est d'évaluer le démêlage entre les caractéristiques d'identité et de fond de teint fait par le réseau.
Pour évaluer la qualité globale du transfert de maquillage, nous avons comparé notre méthode avec des méthodes de l'état de l'art selon trois axes : la conservation de l'identité (IC), la qualité de l'image (FID [42]) et la préservation des détails de l'image (MSE_o). La conservation de l'identité (IC) est quantifiée en calculant la distance entre les caractéristiques intermédiaires du réseau de reconnaissance faciale FaceNet [43]. La préservation des détails de l'image (MSE_o) est mesurée en calculant la distance pixel à pixel. Dans cette étude, les caractéristiques de l'arrière-plan, des cheveux et des yeux de l'image source ont été comparées à celles de l'image générée.

4.2 Résultat Qualitatif

Les résultats sont analysés quantitativement selon trois axes : le teint de la peau, la conservation des zones non maquillées (arrière-plan, cheveux et yeux) et le processus de démaquillage. La figure 4 présente des exemples obtenus via notre modèle. Les architectures de l'état de l'art reproduisent à la fois le maquillage, la lumière et le teint de la référence sur la source. Cette observation est vraie à la fois pour la tâche de transfert de maquillage (première, deuxième et troisième lignes) et de démaquillage (sixième ligne). Cela est dû à la technique de correspondance d'histogramme, qui prend la couleur (à la fois les couleurs de peau, de lumière et de maquillage) de la référence et l'applique au visage de la source, ce qui conduit à un résultat moins réaliste. Notre architecture génère une nouvelle couleur pour le teint de la peau qui semble plus plausible. Nous observons également, principalement pour SSAT [5] et EleGANt [1], des changements dans les zones non maquillées, tels que des changements de couleur d'arrière-plan (dans la quatrième ligne) et des changements de couleur de cheveux (dans les première et troisième lignes). Ces deux changements sont illustrés dans la figure 5. Grâce à la fonction de perte *Makeup Free Area Loss*, nous réduisons les changements dans les zones non maquillées telles que les cheveux ou l'arrière-plan. Certains artefacts sont visibles dans les images générées par SSAT. Ils s'expliquent par un cadrage différent entre les images apprises et les images d'inférence [44].
La figure 6 compare les images générées par notre méthode et les méthodes de l'état de l'art pour la même source et plusieurs références. SSAT [5] et SpMT [6] produisent un maquillage irréaliste qui modifie complètement le teint de la peau. Les nuances de fond de teint d'EleGANt [1] ne sont pas naturelles, alors que notre modèle génère un maquillage réaliste sur les images générées. Enfin, notre modèle effectue moins de changements entre les images générées que les modèles de l'état de l'art. Ceci est également confirmé par des résultats quantitatif (en utilisant la mé-

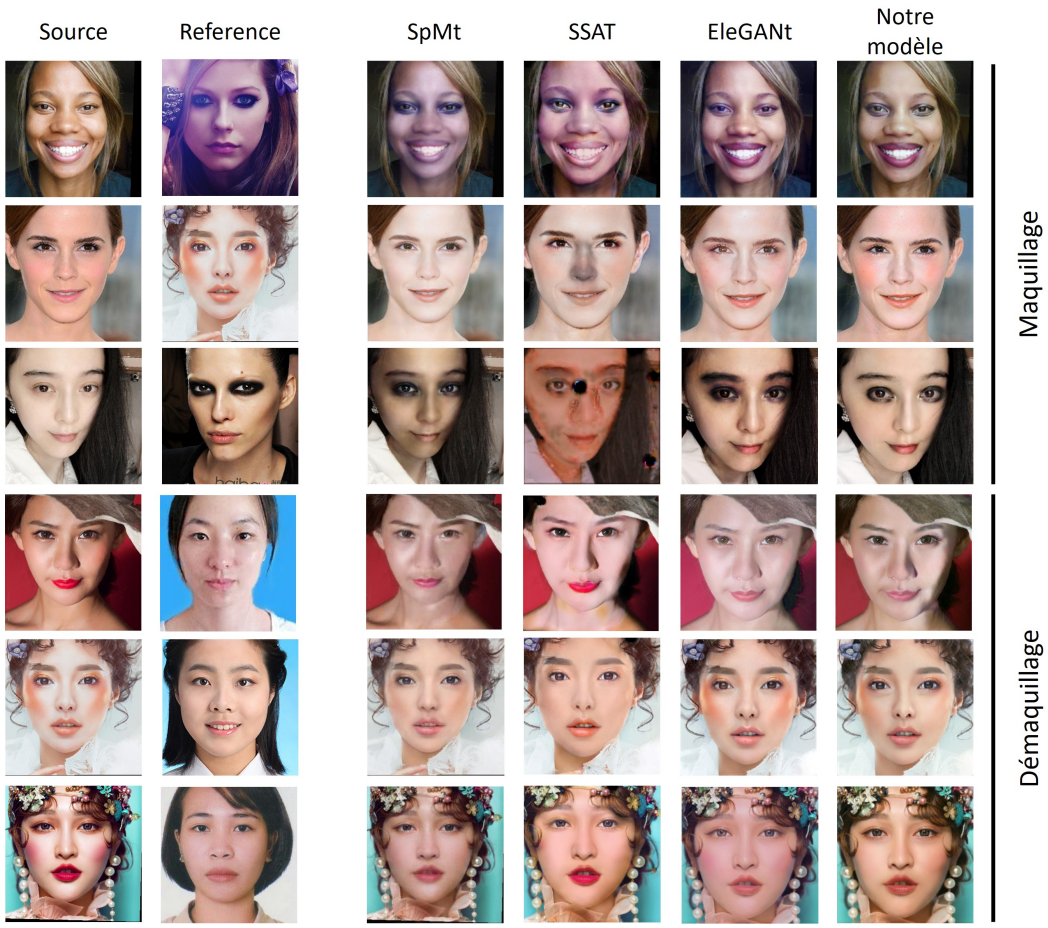

FIGURE 4 – Comparaison avec les méthodes SpMT[6], SSAT[5], EleGANt[1]. Les deux premières lignes montrent des exemples dont les teints de peau sont très différents entre la référence et la source. Contrairement aux autres méthodes qui transfèrent le teint de la peau et le maquillage, le teint de la source est préservé lors de l'application du maquillage. Les troisième et quatrième lignes montrent que notre méthode conserve les zones non maquillées telles que les cheveux ou l'arrière-plan. Les trois dernières lignes montrent la tâche de démaquillage. Notre modèle produit un teint de peau plus naturel que les autres architectures de l'état de l'art. Les images présentées proviennent de l'ensemble de données MT.

FIGURE 5 – Zoom sur les détails de la comparaison avec les méthodes SpMT[6], SSAT[5], EleGANt[1]. La première et deuxième lignes montrent, respectivement, les changements de couleur des cheveux et de couleur de l'arrière-plan dans la troisième et quatrième lignes de la figure 4.

	$FDT\ Er. \downarrow$	$Chc\ Er. \downarrow$	$MSE_o \downarrow$
EleGANt [1]	2.84	0.33	8.42
+ Makeup-Free loss	2.83	0.36	8.26
+ Double décodeur	2.70	0.31	8.35
+ Amélioration PGT	2.77	0.23	8.29

TABLE 1 – Impact de la contribution proposée sur la précision du fond de teint ($FDT\ Er.$), l'erreur de cohérence ($Chc\ Er.$) et l'erreur quadratique moyenne sur l'arrière-plan et les cheveux (MSE_o)

trique de l'erreur de cohérence).

4.3 Impact individuel des contributions

Dans cette section, nous démontrons l'impact de chacune des trois contributions sur les résultats, en utilisant la Figure 7 et le Tableau 1. Les deux premières colonnes de la Figure 7 représentent respectivement la source et la référence. La troisième colonne montre l'épine dorsale EleGANt seule comme ligne de base.

La première contribution est la fonction de perte *Makeup Free Area Loss* (quatrième colonne de la Figure 7). Comme

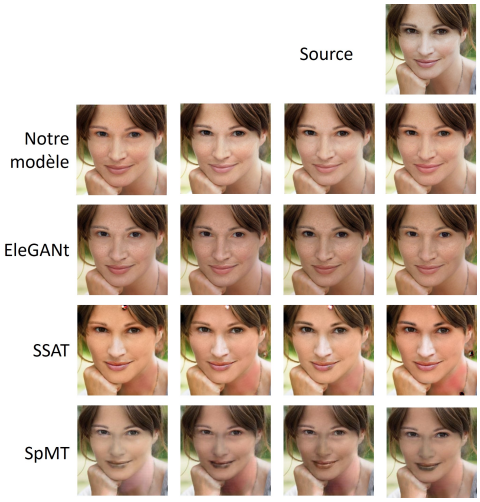

FIGURE 6 – Comparaison des images générées pour **l'erreur de cohérence** avec l'état de l'art [1] [5] [6]. L'image source provient du MTDataset et les références sont issues de la base de données de vérité terrain. Chaque référence a la même composition.

le montrent les deuxième et quatrième lignes de la Figure 7, la couleur des cheveux change dans la base EleGANt tandis qu'avec la *Makeup Free Area Loss*, les cheveux restent inchangés. Ceci est confirmé par la métrique MSE_o dans le tableau 1.

Notre deuxième contribution est l'architecture à double décodeur (cinquième colonne de la figure 7). Cette contribution vise à obtenir un rendu plus réaliste en séparant les tâches de maquillage et de démaquillage. Cette architecture aide le GAN à démêler les caractéristiques d'identité et de maquillage. Dans la Figure 7, cette contribution aide à préserver le blush dans le processus de maquillage (première ligne) et à le supprimer dans le processus de démaquillage (troisième ligne). Le double décodeur permet de supprimer complètement le contour des yeux dans la quatrième ligne. Quantitativement, le double décodeur réduit l'erreur de précision et de cohérence du fond de teint. Cela suggère un meilleur démêlage entre le fond de teint et la couleur de la peau.

Notre dernière contribution est l'amélioration du PGT via le nouveau transfert de fond de teint. Son objectif est également de préserver autant que possible l'identité de la source en ne transférant que la couleur du fond de teint et non la couleur de peau de la référence. Ceci est illustré dans la deuxième ligne de la Figure 7, où nous montrons un cas extrême où la source et la référence ont deux teints de peau différents. Nous observons que seule l'amélioration du PGT permet de préserver le teint original de la source et réduit l'erreur de cohérence du fond de teint (tableau 1), ce qui suggère une meilleure extraction et représentation des caractéristiques du fond de teint.

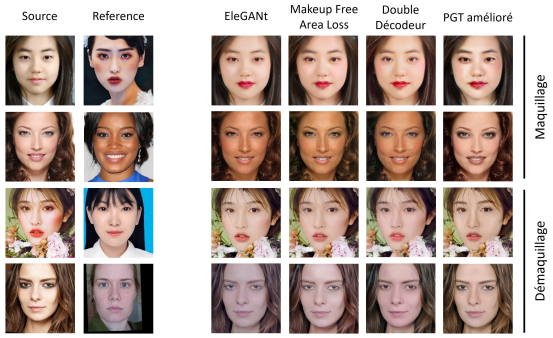

FIGURE 7 – L'étude d'ablation sur les tâches de maquillage et de démaquillage. La première colonne montre la ligne de base EleGANt [1]. La deuxième colonne montre la perte de surface sans maquillage. La troisième colonne ajoute l'architecture du double décodeur et la dernière colonne montre le processus de génération de α.

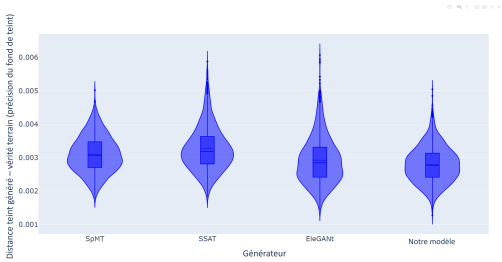

FIGURE 8 – Comparaison de la **précision du fond de teint** avec l'état de l'art [1] [5] [6]. La précision du fond de teint est mesurée par la distance entre les images générées et leur vérité terrain. Ce violin plot a été généré à l'aide de la base de données privée avec vérité terrain.

4.4 Résultat Quantitatif

Métrique avec vérité terrain Pour analyser quantitativement les résultats, deux mesures sont utilisées : la précision du fond de teint et l'erreur de cohérence (décrites au point 4.1). Ces mesures sont appliquées aux images de la base de données privée contenant les vérités terrain. La figure 8 montre la précision du fond de teint des méthodes de l'état de l'art et de notre méthode. On observe une meilleure proximité avec la réalité qui s'explique par la capacité de notre méthode à garder l'information du teint de la source au lieu de copier toutes les couleurs. Notre modèle a la meilleure précision du fond de teint (notre modèle a une erreur médiane de **0.00277** alors que SpMT [6], SSAT [5], EleGANt [1] ont une erreur médiane respectivement de 0.00306, 0.00318, et 0.00284). Notre modèle présente également un écart-type de la précision du fond de teint inférieur à celui des autres modèles.

Nous mesurons maintenant l'erreur de cohérence du transfert de fond de teint. La figure 9 montre que les images générées par notre architecture sont plus cohérentes : elles présentent moins d'écarts pour la même source et le même

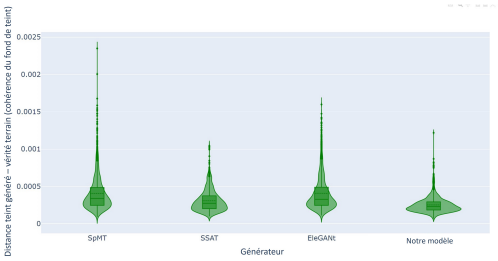

FIGURE 9 – Comparaison de **l'erreur de cohérence** avec l'état de l'art [1] [5] [6]. L'erreur de cohérence mesure l'écart entre les images générées pour la même source et le même maquillage (mais avec des personnes différentes).

	$IC \downarrow$	$FID \downarrow$	$MSE_o \downarrow$
SpMT [6]	0.29	107	**8.24**
SSAT [5]	0.33	117	9.10
EleGANt [1]	**0.24**	<u>84</u>	8.42
Notre modèle	<u>0.25</u>	**80**	<u>8.29</u>

TABLE 2 – Comparaison de la conservation de l'identité (IC), de l'erreur quadratique moyenne sur le fond et les cheveux (MSE_o) et du FID [42]. Les résultats du MSE_o sont arrondis à $1e^{-3}$.

fond de teint (mais avec différents porteurs). Ce résultat suggère que notre architecture démêle mieux les caractéristiques du maquillage et de l'identité grâce à notre architecture à double décodeur couplée au générateur α. Avec une médiane de **0.00023**, notre méthode présente le meilleur résultat en matière d'erreur de cohérence. En comparaison, la médiane est de 0.00034 pour SpMT [6], 0.00027 pour SSAT [5] et 0.00033 pour EleGANt [1]. Notre modèle est également plus constant (i.e., un écart-type plus faible) que les méthodes de l'état de l'art [1] [5] [6].
Nos deux mesures fournissent deux informations : La précision du fond de teint montre la proximité de l'image générée avec sa vérité terrain, et l'erreur de cohérence montre la capacité du modèle à démêler les caractéristiques d'identité et de maquillage. Notre méthode surpasse les précédentes méthodes de transfert de maquillage sur ces deux métriques.

Métrique sans vérité terrain Les résultats du tableau 2 montrent que la méthode proposée est comparable aux méthodes de l'état de l'art en termes de conservation de l'identité (IC), de qualité d'image (FID) et de préservation des détails de l'image (MSE_o). Plus précisément, notre méthode obtient des résultats similaires à EleGANt [1] pour la conservation de l'identité (IC) et la qualité de l'image (FID) tout en surpassant SSAT [5] et SpMT [6]. Pour la conservation des détails de l'image (MSE_o), l'approche proposée obtient des résultats équivalents à ceux de SpMT [6] et surpasse SSAT [5] et EleGANt [1]. Cela montre l'effet positif de la *Makeup Free Area Loss* (3.5). Pour SpMT [6], ce résultat s'explique par le démêlage résultant du module non paramétrique Semantic Aware Correspondence.

	$FDT \uparrow$	$Identit \uparrow$	$Globale \uparrow$
SpMT [6]	11	11	9
SSAT [5]	13	10	10
EleGANt [1]	<u>23</u>	<u>26</u>	<u>31</u>
Notre modèle	**53**	**52**	**51**

TABLE 3 – Résultats de l'étude utilisateurs (ratio des meilleurs résultats) pour le réalisme du transfert de fond de teint (FDT), la préservation de l'identité ($Identit$) et la qualité globale du transfert du maquillage ($Globale$).

Etude utilisateurs Nous avons mené une étude auprès d'utilisateurs afin d'évaluer notre méthode par rapport aux méthodes de transfert de maquillage de l'état de l'art : SpMT [6], SSAT [5], et EleGANt [1]. Nous avons sélectionné au hasard 20 images sources non maquillées et 20 images de référence maquillées afin d'obtenir 400 images générées (pour la tâche de maquillage uniquement) pour chaque méthode. 40 participants ont été invités à choisir la meilleure image en fonction de trois aspects : le réalisme du transfert du fond de teint/blush, la préservation de l'identité (préservation des caractéristiques de l'identité telles que la couleur des yeux et des cheveux) et la qualité globale du transfert du maquillage (en tenant compte de la qualité visuelle, de la fidélité du maquillage transféré, etc.) Pour chaque participant, l'ordre de placement des 4 images synthétiques est aléatoire et les participants sont invités à choisir leur résultat préféré. Au total, nous avons recueilli 568 votes et calculé le pourcentage de votes pour chaque méthode. Les résultats sont présentés dans le tableau 3. Notre méthode reçoit le pourcentage de votes le plus élevé pour chaque aspect, montrant un meilleur transfert du fond de teint, une meilleure préservation de l'identité et un meilleur transfert de maquillage dans l'ensemble.

5 Conclusion

Le transfert de maquillage ne peut être réduit à une copie de couleur. Pour obtenir un meilleur transfert de maquillage de fond de teint, nous avons amélioré l'idée de PGT en remplaçant la technique de correspondance d'histogramme par une méthode d'α blending, dont les paramètres sont appris de manière auto-supervisée. Nous avons introduit une architecture avec deux décodeurs pour éviter l'emmêlement des processus de maquillage et de démaquillage. Enfin, la *Makeup Free Area Loss* aide le générateur à conserver les zones non maquillées (l'arrière-plan, les vêtements, le cou, etc.) inchangées.

Références

[1] C. Yang, W. He, Y. Xu, and Y. Gao, *EleGANt : Exquisite and Locally Editable GAN for Makeup Transfer*, vol. 13676 of *Lecture notes in computer science*, p. 737–754. Springer Nature Switzerland, 2022.

[2] F. He, K. Bai, Y. Zong, Y. Zhou, Y. Jing, G. Wu, and C. Wang, "Makeup transfer : A review," *IET Computer Vision*, Oct 2022.

[3] F. Zhang and C. Yan, "Development and application of facial makeup transfer," in *2021 17th International Conference on Computational Intelligence and Security (CIS)*, p. 304–308, IEEE, Nov 2021.

[4] X. Ou, S. Liu, X. Cao, and H. Ling, *Beauty eMakeup : A Deep Makeup Transfer System*, p. 701–702. ACM Press, Oct 2016.

[5] Z. Sun, Y. Chen, and S. Xiong, "Ssat : A symmetric semantic-aware transformer network for makeup transfer and removal," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, p. 2325–2334, Jun 2022.

[6] M. Zhu, Y. Yi, N. Wang, X. Wang, and X. Gao, "Semi-parametric makeup transfer via semantic-aware correspondence," *arXiv*, 2022.

[7] X. Zhong, X. Huang, Z. Wu, G. Lin, and Q. Wu, "Sara : Controllable makeup transfer with spatial alignment and region-adaptive normalization," *arXiv*, 2023.

[8] S. Liu, X. Ou, R. Qian, W. Wang, and X. Cao, "Makeup like a superstar : Deep localized makeup transfer network," *arXiv*, 2016.

[9] P. Maitra, A. Balina, S. Carlo, and J. R. Glynn, "Optical tools to assess naturalness of cosmetic films," *Color Research and Application*, vol. 34, p. 170–172, Apr 2009.

[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *arXiv*, 2014.

[11] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *arXiv*, 2018.

[12] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 2332–2341, IEEE, Jun 2019.

[13] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *arXiv*, 2018.

[14] R. Abdal, Y. Qin, and P. Wonka, "Image2stylegan : How to embed images into the stylegan latent space ?," *arXiv*, 2019.

[15] Y. Choi, M. Choi, M. Kim, J.-W. SunghunKim, and J. Choo, "Stargan : Unified generative adversarial networks for multi-domain image-to-image translation," *Arxiv*, 2017.

[16] T. Li, R. Qian, C. Dong, S. Liu, Q. Yan, W. Zhu, and L. Lin, "Beautygan : Instance-level facial makeup transfer with deep generative adversarial network," in *Proceedings of the 26th ACM international conference on Multimedia*, p. 645–653, ACM, Oct 2018.

[17] H.-J. Chen, K.-M. Hui, S.-Y. Wang, L.-W. Tsao, H.-H. Shuai, and W.-H. Cheng, "Beautyglow : On-demand makeup transfer framework with reversible generative network," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 10034–10042, IEEE, Jun 2019.

[18] H. Deng, C. Han, H. Cai, G. Han, and S. He, "Spatially-invariant style-codes controlled makeup transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6549–6557, June 2021.

[19] W. Jiang, S. Liu, C. Gao, J. Cao, R. He, J. Feng, and S. Yan, "Psgan : Pose and expression robust spatial-aware gan for customizable makeup transfer," *arXiv*, 2019.

[20] S. Liu, W. Jiang, C. Gao, R. He, J. Feng, B. Li, and S. Yan, "Psgan++ : Robust detail-preserving makeup transfer and removal.," *IEEE transactions on pattern analysis and machine intelligence*, vol. PP, May 2021.

[21] T. Nguyen, A. T. Tran, and M. Hoai, "Lipstick ain't enough : Beyond color matching for in-the-wild makeup transfer," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 13300–13309, IEEE, Jun 2021.

[22] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, p. 2242–2251, IEEE, Oct 2017.

[23] A. Dhall, G. Sharma, R. Bhatt, and G. M. Khan, *Adaptive Digital Makeup*, vol. 5876 of *Lecture notes in computer science*, p. 728–736. Springer Berlin Heidelberg, 2009.

[24] Y. Zhang, H. Cui, Y. Li, and Z. Feng, "Makeup based on segmentation and local transfer," in *2019 6th International Conference on Behavioral, Economic and Socio-Cultural Computing (BESC)*, p. 1–6, IEEE, Oct 2019.

[25] K. Scherbaum, T. Ritschel, M. Hullin, T. Thormählen, V. Blanz, and H.-P. Seidel, "Computer-suggested facial makeup," *Computer Graphics Forum*, vol. 30, p. 485–492, Apr 2011.

[26] J. Xiang, J. Chen, W. Liu, X. Hou, and L. Shen, *RamGAN : Region Attentive Morphing GAN for Region-Level Makeup Transfer*, vol. 13682 of *Lecture notes in computer science*, p. 719–735. Springer Nature Switzerland, 2022.

[27] Y. Lyu, P. Chen, J. Sun, B. Peng, X. Wang, and J. Dong, "Dran : Detailed region-adaptive normalization for conditional image synthesis," *arXiv*, 2021.

[28] R. Kips, P. Gori, M. Perrot, and I. Bloch, *CA-GAN : Weakly Supervised Color Aware GAN for Controllable Makeup Transfer*, vol. 12537 of *Lecture notes in computer science*, p. 280–296. Springer International Publishing, 2020.

[29] Q. Gu, G. Wang, M. T. Chiu, Y.-W. Tai, and C.-K. Tang, *LADN : Local Adversarial Disentangling Network for Facial Makeup and De-Makeup*, p. 10480–10489. IEEE, Oct 2019.

[30] H. Chang, J. Lu, F. Yu, and A. Finkelstein, *Pairedcyclegan : asymmetric style transfer for applying and removing makeup*, p. 40–48. IEEE, Jun 2018.

[31] S. Hu, X. Liu, Y. Zhang, M. Li, L. Y. Zhang, H. Jin, and L. Wu, *Protecting Facial Privacy : Generating Adversarial Identity Masks via Style-robust Makeup Transfer*, p. 14994–15003. IEEE, Jun 2022.

[32] H. Kazemi, S. M. Iranmanesh, and N. Nasrabadi, "Style and content disentanglement in generative adversarial networks," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, p. 848–856, IEEE, Jan 2019.

[33] H. Zhang, W. Chen, H. He, and Y. Jin, "Disentangled makeup transfer with generative adversarial network," *arXiv*, 2019.

[34] Z. Sun, F. Liu, W. Liu, S. Xiong, and W. Liu, *Local facial makeup transfer via disentangled representation*, vol. 12625 of *Lecture notes in computer science*, p. 459–473. Springer International Publishing, 2021.

[35] A. Neumann and L. Neumann, "Color style transfer techniques using hue, lightness and saturation histogram matching," *The Eurographics Association*, 2005.

[36] T. Porter and T. Duff, "Compositing digital images," in *Proceedings of the 11th annual conference on Computer graphics and interactive techniques - SIGGRAPH '84* (H. Christiansen, ed.), p. 253–259, ACM Press, 1984.

[37] B. Evangelista, H. Meshkin, H. Kim, A. Aburto, B. M. Rubinstein, and A. Ho, "Realistic ar makeup over diverse skin tones on mobile," in *SIGGRAPH Asia 2018 Posters on - SA '18*, p. 1–2, ACM Press, Dec 2018.

[38] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "Bisenet v2 : Bilateral network with guided aggregation for real-time semantic segmentation," *International journal of computer vision*, vol. 129, p. 3051–3068, Nov 2021.

[39] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan : towards diverse and interactive facial image manipulation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 5548–5557, IEEE, Jun 2020.

[40] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Springer Science and Business Media LLC*, 2000.

[41] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," *arXiv*, 2017.

[42] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *arXiv*, 2017.

[43] F. Schroff, D. Kalenichenko, and J. Philbin, *FaceNet : A unified embedding for face recognition and clustering*, p. 815–823. IEEE, Jun 2015.

[44] Y. Wang and Z. Sun, "Generate artifacts that appear in the final image," Jul 2022.