Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων

Εργαστηριο Τεχνητης Νοημοσυνης και Συστηματων Μαθησης

Language-based Interpretation of Generative Models

Diploma Thesis

by

Aristotelis Koutris

Επιβλέπων: Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2023

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Πληροφορικής
Εργαστήριο Τεχνητής Νοημοσύνης και Συστημάτων Μάθησης

Language-based Interpretation of Generative Models

Diploma Thesis

by

Aristotelis Koutris

Επιβλέπων: Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 13η Μαρτίου, 2023.

.....	.....	.....
Γεώργιος Στάμου	Αθανάσιος Βουλόδημος	Στέφανος Κόλλιας
Καθηγητής Ε.Μ.Π.	Επ. Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2023

.....

Αριστοτελης Κουτρης
Διπλωματούχος Ηλεκτρολόγος Μηχανικός
και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Τα γεννητικά μοντέλα έχουν επιδείξει σημαντική πρόοδο στη δημιουργία ρεαλιστικών εικόνων και χρησιμοποιούνται όλο και περισσότερο σε μια ποικιλία εφαρμογών. Ωστόσο, η ερμηνεία και η κατανόηση αυτών των μοντέλων παραμένει μια πρόκληση. Η εργασία αυτή πραγματεύεται δύο κύρια θέματα για την αντιμετώπιση του προβλήματος αυτού.

Το πρώτο θέμα επικεντρώνεται στον σχεδιασμό μια μεθόδου για την ανακάλυψη ερμηνεύσιμων κατευθύνσεων στον λανθάνοντα χώρο του Glow. Το Glow είναι ένα γεννητικό μοντέλο ροής το οποίο διαθέτει έναν υψηλά αποσυσχετισμένο λανθάνοντα χώρο, ο οποίος κωδικοποιεί τα σημασιολογικά χαρακτηριστικά μιας εικόνας σε ανεξάρτητές λανθάνουσες μεταβλητές. Η ιδιότητα του αυτή, σε συνδυασμό με το ότι η αρχιτεκτονική του είναι αντιστρέψιμη, το καθιστούν ένα πολύ χρήσιμο μοντέλο για την τροποποίηση εικόνων μέσω της λανθάνουσας αναπαράστασης τους. Η μέθοδος που προτείνουμε, επιτρέπει την εύρεση λανθανουσών κατευθύνσεων που αντιστοιχούν σε ένα σημασιολογικό χαρακτηριστικό της εικόνας, με βάση μία κειμενική περιγραφή που το περιγράφει. Η καθοδήγηση της μεθόδου από φυσική γλώσσα της δίνει μεγαλύτερη ευελιξία σε σχέση με άλλες επιβλεπόμενες ή μη επιβλεπόμενες μεθόδους εξερεύνησης του λανθάνοντος χώρου με τις οποίες συγκρίνουμε τα αποτελέσματα μας.

Με αφορμή την μεγάλη άνοδο των μοντέλων σύνθεσης εικόνας από κείμενο και την αποτελεσματικότητα των μοντέλων διάχυσης στον τομέα αυτό, προτείνουμε επίσης, μια μέθοδο για την συστηματική αξιολόγηση του Stable Diffusion. Πιο συγκεκριμένα, συνθέτοτουμε εικόνες από ένα σύνολο ιεραρχικά συνδεδεμένων εννοιών του WordNet, και εξετάζουμε σε ποιό βαθμό η ιεραρχία των εννοιών αποτυπώνεται στις κατανομές εικόνων που συνθέσαμε. Με τον τρόπο αυτό ποσοτικοποιούμε την δυνατότητα του Stable Diffusion να διαφοροποιεί μεταξύ στενά συνδεδεμένων εννοιών, και ανιχνεύουμε τυχόν προκαταλήψεις που υπάρχουν υπέρ συγκεκριμένων εννοιών.

Λέξεις-κλειδιά — Σύνθεση Εικόνας απο Κείμενο, Λανθάνων Χώρος, Χειρισμός Εικόνων, Γεννητικά Μοντέλα Ροής, Μοντέλα Διάχυσης

Abstract

Generative models have shown remarkable progress in generating realistic images and are being increasingly used in a variety of applications. However, interpreting and understanding these models remains a challenge. Two main topics have been addressed in this thesis to tackle this problem.

The first topic focuses on Glow, a flow-based generative model with exact latent-variable inference and log-likelihood. The key advantages of Glow are its invertibility and the ability to perform easy image manipulation through its latent space. This thesis proposes a novel framework for interpretable latent direction discovery in the latent space of Glow, by leveraging the text-guided image generation and manipulation capabilities of StyleCLIP. The framework is compared with existing state-of-the-art supervised and unsupervised latent direction discovery methods.

Secondly, motivated by the rapid growth of text-guided image generation and the effectiveness of diffusion models such as Stable Diffusion, this thesis proposes a systematic method to evaluate Stable Diffusion's ability to model and generate images from closely related concepts using WordNet. This study enables the detection of potential biases towards different areas of the distribution modelled by the generative model.

Overall, this thesis aims to provide a better understanding of generative models by proposing novel frameworks and evaluation methodologies for their interpretability and effectiveness. These contributions can have important implications for improving the applicability and reliability of generative models in various fields.

Keywords — Text-Guided Image Generation, Latent Space, Image Manipulation, Flow-based Generative Models, Diffusion Models

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου, κ. Γιώργο Στάμου για την ευκαιρία που μου έδωσε να εκπονήσω την διπλωματική αυτή στο Εργαστήριο Τεχνητής Νοημοσύνης και Συστημάτων Μάθησης. Επίσης, θα ήθελα να ευχαριστήσω τον κ. Σιόλα για την πρόσβαση σε υπολογιστικούς πόρους και την βοήθεια που μου προσέφερε.

Ευχαριστώ ιδιαίτερα και την Μαρία Λυμπεραίου για την πολύτιμη συνεισφορά της και την καθοδήγηση που μου παρείχε καθόλη την διάρκεια εκπόνησης αυτής της διπλωματικής εργασίας.

Με αφορμή την ολοκλήρωση των σπουδών μου, θα ήθελα επίσης να ευχαριστήσω την οικογένεια μου, και πρωτίστως τους γονείς μου, για την υποστήριξη και την πίστη μου έχουν δείξει σε μένα. Τέλος, θα ήθελα να ευχαριστήσω τους φίλους, συμφοιτητές και μη, με τους οποίος μοιράστηκα τα φοιτητικά μου χρόνια.

Αριστοτέλης Κούτρης, Μάρτιος 2023

Contents

List of Figures

Chapter 1

Εκτεταμένη Περίληψη στα Ελληνικά

1.1 Θεωρητικό Υπόβαθρο

Η τεχνητή νοημοσύνη έχει φέρει επανάσταση στον τρόπο με τον οποίο αλληλεπιδρούμε με την τεχνολογία και έχει σημειώσει αξιοσημείωτη πρόοδο σε διάφορες εφαρμογές, όπως η επεξεργασία φυσικής γλώσσας, η όραση υπολογιστών και η ρομποτική. Μεταξύ αυτών των εφαρμογών, η δημιουργία εικόνων είναι ένας ταχέως εξελισσόμενος ερευνητικός τομέας που έχει σημειώσει σημαντική πρόοδο τα τελευταία χρόνια. Η παραγωγή εικόνων αναφέρεται στη διαδικασία δημιουργίας ρεαλιστικών εικόνων από το μηδέν ή τροποποίησης υφιστάμενων εικόνων με τη χρήση αλγορίθμων τεχνητής νοημοσύνης. Η χρήση της τεχνητής νοημοσύνης στη δημιουργία εικόνων έχει οδηγήσει στην ανάπτυξη ισχυρών εργαλείων για καλλιτέχνες, σχεδιαστές και κινηματογραφιστές που δημιουργούν με ευκολία οπτικές εικόνες υψηλής ποιότητας. Επιπλέον, οι τεχνικές αυτές έχουν βρει εφαρμογές σε διάφορους τομείς, όπως η ιατρική, η ψυχαγωγία και τα παιχνίδια. Για παράδειγμα, οι ιατρικές εικόνες που παράγονται με τεχνητή νοημοσύνη μπορούν να βοηθήσουν τους γιατρούς να διαγνώσουν με ακρίβεια τις ασθένειες, ενώ οι εικόνες που παράγονται με τεχνητή νοημοσύνη στις βιομηχανίες παιχνιδιών και ψυχαγωγίας μπορούν να προσφέρουν μια πιο καθηλωτική εμπειρία στους χρήστες.

Οι εξελίξεις στην παραγωγή εικόνων έχουν οδηγήσει στην ανάπτυξη πολλών εξελιγμένων μοντέλων, όπως τα Αναγεννητικά Ανταγωνιστικά Δίκτυα (Generative Adversarial Networks ή GANs) [1], οι Εναλασσόμενοι Αυτοκωδικοποιητές (Variational Autoencoders ή VAEs) [2] και τα μοντέλα διάχυσης (Diffusion Models ή DMs) [3]. Αυτά τα μοντέλα μπορούν να παράγουν εικόνες υψηλής ποιότητας που είναι σχεδόν δυσδιάκριτες από τις πραγματικές εικόνες. Επιπλέον, μπορούν να εκπαιδευτούν σε ένα μεγάλο σύνολο δεδομένων εικόνων για να μάθουν και να μιμηθούν το ύφος ενός συγκεκριμένου καλλιτέχνη, εποχής ή είδους. Πιο πρόσφατα, η εμφάνιση μοντέλων δημιουργίας εικόνων με τη βοήθεια κειμένου επέτρεψε τη δημιουργία και επεξεργασία εικόνων από περιγραφές φυσικής γλώσσας, παρέχοντας ένα πιο διαισθητικό και προσιτό μέσο για τη δημιουργία οπτικού περιεχομένου [4, 5, 6, 7].

Η κατανόηση του τρόπου με τον οποίο τα γεννητικά μοντέλα αναπαριστούν διαφορετικές έννοιες είναι ζωτικής σημασίας για διάφορους λόγους. Πολλοί τύποι γεννητικών μοντέλων μαθαίνουν να αναπαριστούν σύνθετες οπτικές έννοιες χρησιμοποιώντας έναν λανθάνοντα χώρο χαμηλής διάστασης, όπου κάθε διάσταση αντιστοιχεί σε ένα διαφορετικό χαρακτηριστικό της εικόνας. Αναλύοντας τον λανθάνοντα χώρο, μπορούμε να αποκτήσουμε γνώσεις σχετικά με τον τρόπο με τον οποίο το μοντέλο αναπαριστά διαφορετικές έννοιες και να χρησιμοποιήσουμε αυτή τη γνώση για να χειριστούμε ή να δημιουργήσουμε εικόνες με συγκεκριμένα χαρακτηριστικά. Μπορούμε επίσης να αποκτήσουμε γνώσεις σχετικά με τους περιορισμούς και τις πιθανές προκαταλήψεις των γεννητικών μοντέλων. Για παράδειγμα, ένα μοντέλο που έχει εκπαιδευτεί σε ένα συγκεκριμένο σύνολο δεδομένων μπορεί να μην γενικεύεται καλά σε νέα δεδομένα ή έννοιες που δεν αντιπροσωπεύονται καλά στα δεδομένα εκπαίδευσης. Ως εκ τούτου, είναι σημαντικό να αναλύουμε την απόδοση του μοντέλου σε διαφορετικές εφαρμογές και να αξιολογούμε τα δυνατά και αδύνατα σημεία του, ώστε να προσδιορίσουμε πού μπορεί να χρειάζεται βελτίωση.

1.1.1 Εξευρεύνηση Λανθάνοντος Χώρου

Ο λανθάνων χώρος (latent space) των γεννητικών δικτύων είναι μια θεμελιώδης έννοια της βαθιάς μάθησης που έχει σημαντικές επιπτώσεις για τον χειρισμό και τη σύνθεση εικόνων. Ο λανθάνων χώρος αναφέρεται στον υψηλής διάστασης χώρο λανθάνουσας μεταβλητής που χρησιμοποιεί ένα γεννητικό δίκτυο για να αντιστοιχίσει από μια προηγούμενη κατανομή σε μια κατανομή δεδομένων ενδιαφέροντος. Αυτές οι λανθάνουσες μεταβλητές ελέγχουν διάφορες πτυχές της παραγόμενης εικόνας, συμπεριλαμβανομένου του χρώματος, της υφής και του σχήματος.

Ένα από τα βασικά πλεονεκτήματα της εργασίας στον λανθάνοντα χώρο είναι η δυνατότητα εκτέλεσης στοχευμένων χειρισμών εικόνας. Χειριζόμενοι τις λανθάνουσες μεταβλητές, είναι δυνατόν να δημιουργηθούν νέες εικόνες που διαφέρουν από τις αρχικές με συγκεκριμένους τρόπους. Για παράδειγμα, τροποποιώντας τις σχετικές λανθάνουσες μεταβλητές, μπορεί κανείς να αλλάξει τη στάση ή την έκφραση ενός προσώπου σε μια εικόνα ή να αλλάξει την υφή ή το χρώμα ενός αντικειμένου. Επιπλέον, κάνοντας παρεμβολή μεταξύ των λανθανουσών μεταλβητών δύο διαφορετικών εικόνων, όπως απεικονίζεται στο σχήμα 1.1.1 μπορούμε να δούμε την αποτελεσματικότητα με την οποία κωδικοποιούνται τα χαρακτηριστικά της εικόνας. Οι χειρισμοί εικόνων με τη χρήση του λανθάνοντος χώρου των γεννητικών δικτύων έχουν ευρείες εφαρμογές, συμπεριλαμβανομένης της επεξεργασίας εικόνων, της επαύξησης δεδομένων και της σύνθεσης δεδομένων. Στην όραση υπολογιστών, τα γεννητικά δίκτυα χρησιμοποιούνται συχνά για την επαύξηση των συνόλων δεδομένων εκπαίδευσης με τη δημιουργία νέων εικόνων που είναι παρόμοιες με τα αρχικά δεδομένα αλλά έχουν συγκεκριμένες τροποποιήσεις, βελτιώνοντας έτσι την ευρωστία και τη γενίκευση των μοντέλων βαθιάς μάθησης.

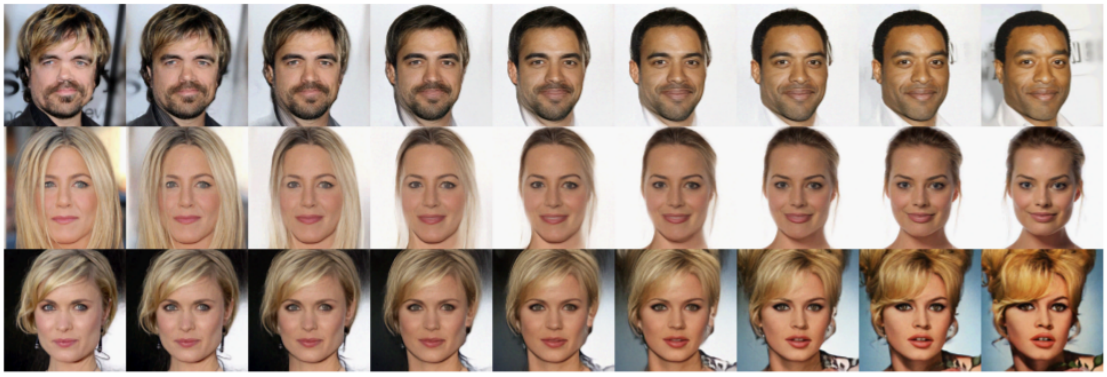


Figure 1.1.1: Γραμμική Παρεμβολή μεταξύ δύο εικόνων στον λανθάνοντα χώρο του Glow [8]

Μια άλλη σημαντική έννοια στον λανθάνοντα χώρο των γεννητικών δικτύων είναι η αποσυσχέτιση (disentanglement). Η αποσυσχέτιση αναφέρεται στην ικανότητα διαχωρισμού των λανθανουσών μεταβλητών σε ανεξάρτητους και ερμηνεύσιμους παράγοντες μεταβολής. Με άλλα λόγια, οι αποσυσχέτισμένες αναπαραστάσεις του λανθάνοντος χώρου επιτρέπουν μεγαλύτερο έλεγχο σε συγκεκριμένες πτυχές της παραγόμενης εικόνας. Για παράδειγμα, μια αποσυσχέτισμένη αναπαράσταση του λανθάνοντος χώρου περιοχής για τα πρόσωπα μπορεί να διαχωρίσει τη στάση και την έκφραση του προσώπου σε διαφορετικούς παράγοντες μεταβολής. Αυτός ο διαχωρισμός θα επέτρεπε μεγαλύτερο έλεγχο του χειρισμού αυτών των συγκεκριμένων χαρακτηριστικών.

1.2 Σχετικές Αρχιτεκτονικές

1.2.1 Glow

Το μοντέλο Glow [8] είναι ένα βαθύ γεννητικό μοντέλο που υλοποιεί μια κανονικοποιητική ροή [9] συσσωρεύοντας μια ακολουθία αντιστρέψιμων συναρτήσεων μετασχηματισμού. Η αρχιτεκτονική του Glow, που βασίζεται στα μοντέλα ροής NICE [10] και RealNVP [11], συγκροτείται από βήματα ροής (flow steps) κάθε ένα εκ των οποίων περιλαμβάνει τρια διακριτά επίπεδα.

Το πρώτο επίπεδο είναι ένα επίπεδο actnorm, το οποίο κανονικοποιεί την έξοδο του προηγούμενου επιπέδου με κλιμάκωση και μετατόπιση ώστε να έχει μηδενική μέση τιμή και μοναδιαία διακύμανση κατά μήκος κάθε καναλιού.

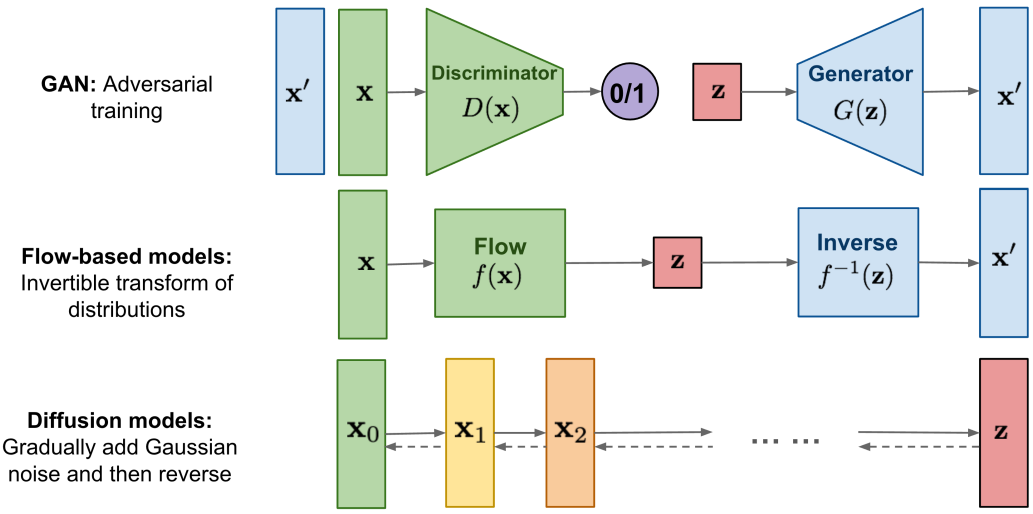


Figure 1.2.1: Επισκόπηση των τριών τύπων γεννητικών μοντέλων που χρησιμοποιούνται στην παρούσα εργασία.

Το δεύτερο επίπεδο είναι ένα αντιστρέψιμο επίπεδο συνέλιξης 1 x 1 που αναμειγνύει τα κανάλια εισόδου για να επιτρέψει τις αλληλεπιδράσεις μεταξύ τους. Αυτό το στρώμα έχει έναν πίνακα βαρών που αρχικοποιείται ως ένας τυχαίος ορθογώνιος πίνακας για να εξασφαλιστεί ότι είναι αντιστρέψιμος.

Το τρίτο επίπεδο είναι ένα επίπεδο σύζευξης, το οποίο χωρίζει τα κανάλια εισόδου σε δύο ομάδες και εφαρμόζει έναν αντιστρέψιμο μετασχηματισμό στη μία ομάδα με βάση την άλλη ομάδα. Αυτό το στρώμα εξασφαλίζει ότι η έξοδος παραμένει αντιστρέψιμη και διατηρεί τις διαστάσεις της εισόδου, ενώ εισάγει μη γραμμικότητα με τοπικό τρόπο.

Το Glow, όπως όλα τα μοντέλα που βασίζονται στη ροή, είναι αντιστρέψιμο και άρα έχει την ικανότητα να εξάγει ακριβή συμπερασμό των λανθανουσών μεταβλητών μιας δεδομένης εικόνας. Αυτή η δυνατότητα μετασχηματισμού από τον χώρο της εικόνας στον λανθάνοντα χώρο, επιτρέπει ακριβείς χειρισμούς και ανακατασκευές εικόνων. Χειριζόμενοι τις λανθάνουσες μεταβλητές, μπορούμε να ελέγξουμε με ακρίβεια την εμφάνιση της εικόνας που προκύπτει. Αυτό καθιστά το μοντέλο Glow ένα ισχυρό εργαλείο για διάφορες εργασίες δημιουργίας και χειρισμού εικόνων. Ένα άλλο πλεονέκτημα του Glow είναι το υψηλό επίπεδο αποσυσχέτισης που χαρακτηρίζει τον λανθάνοντα χώρο του. Η ικανότητά του να διαχωρίζει τους παράγοντες μεταβολής στον λανθάνοντα χώρο, επιτρέπει τον προσδιορισμό ερμηνεύσιμων κατευθύνσεων που αντιστοιχούν σε συγκεκριμένα σημασιολογικά χαρακτηριστικά της εικόνας. Αυτές οι κατευθύνσεις αντιστοιχούν σε αλλαγές σε συγκεκριμένα σημασιολογικά χαρακτηριστικά των παραγόμενων εικόνων. Μια απο τις κύριες συνεισφορές της εργασίας αυτής είναι η δημιουργία μιας μεθόδου εύρεσης ερμηνεύσιμων κατευθύνσεων του λανθάνοντος χώρου με την βοήθεια γλωσσικών περιγραφών.

1.2.2 StyleCLIP

Το StyleCLIP [12] χρησιμοποιεί έναν συνδυασμό γλωσσικών αναπαραστάσεων και αναπαραστάσεων εικόνας για να καθοδηγήσει τη διαδικασία δημιουργίας εικόνων. Συγκεκριμένα, το StyleCLIP συνδυάζει τη δύναμη του μοντέλου Contrastive Language-Image Pre-training ή CLIP [13] με το StyleGAN 2 [14], για τη δημιουργία εικόνων που είναι τόσο σημασιολογικά όσο και οπτικά συνεπείς με μια δεδομένη κειμενική περιγραφή.

Ένα από τα βασικά πλεονεκτήματα του StyleCLIP είναι η ικανότητά του να παράγει εικόνες και να κάνει τροποποιήσεις που είναι πολύ συγκεκριμένες για τη δεδομένη περιγραφή κειμένου. Αυτό επιτυγχάνεται χάρη στο προεκπαιδευμένο γλωσσικό μοντέλο CLIP, το οποίο έχει εκπαιδευτεί σε ζευγάρια εικόνων και κειμενικών περιγραφών. Το CLIP επιτρέπει την κωδικοποίηση κειμένου και εικόνας σε έναν κοινό χώρο αναπαράστασης (embedding space), δίνοντας την δυνατότητα υπολογισμού της ομοιότητας μεταξύ τους.

Οι δημιουργοί του StyleCLIP περιγράφουν τρεις διαφορετικές προσεγγίσεις που αξιοποιούν τις δυνατότητες του CLIP για την εφαρμογή τροποποίησεων σε μια εικόνα με την χρήση του StyleGAN. Οι προσεγγίσεις αυτές

διαφέρουν ως προς την ταχύτητα, την ερμηνευσιμότητα και την ευελιξία τους. Στην συνέχεια περιγράφουμε τον αλγόριθμο εύρεσης παγκοσμίων κατευθύνσεων (Global Directions) που χρησιμοποιήσαμε για τους χειρισμούς εικόνων στο πλαίσιο αυτής της εργασίας.

Η μέθοδος εύρεσης παγκοσμίων κατευθύνσεων, περιλαμβάνει την εκμάθηση μιας απεικόνισης από το χώρο αναπαράστασης του CLIP στον χώρο style space του StyleGAN. Πιο συγκεκριμένα, η προσέγγιση αυτή χρησιμοποιεί μία κειμενική περιγραφή αναφοράς και μία κειμενική περιγραφή στόχου για να καθορίσει την συνεισφορά κάθε καναλιού του χώρου style space στην επιθυμητή κατεύθυνση χειρισμού. Η μέθοδος ξεκινά με τον υπολογισμό της κατεύθυνση CLIP που αντιστοιχεί στη διαφορά μεταξύ του κειμένου-αναφοράς και του κειμένου-στόχου. Στη συνέχεια, η κατεύθυνση CLIP συγκρίνεται με τις κατευθύνσεις που αντιστοιχούν σε διαταραχές σε κάθε κανάλι του χώρου style space. Κανάλια με συνεισφορά μικρότερη από ένα επιλεγμένο κατώφλι αγνοούνται και μόνο τα κανάλια με υψηλή συνεισφορά χρησιμοποιούνται για τον προσδιορισμό της κατεύθυνσης Δs στο χώρο style space. Αυτή η κατεύθυνση μπορεί στη συνέχεια να χρησιμοποιηθεί για να τροποποιηθεί η παραγόμενη εικόνα ως προς την κατεύθυνση της κειμενικής περιγραφής στόχου. Αυτή η προσέγγιση επιτρέπει τον γρήγορο χειρισμό εικόνων με βάση ενός ζεύγους κειμενικών περιγραφών, χωρίς την ανάγκη βελτιστοποίησης για κάθε διαφορετική εικόνα ή εκπαίδευσης για κάθε διαφορετική κειμενική περιγραφή όπως οι δύο άλλες μέθοδοι. Επιπλέον, η μέθοδος αυτή επιτρέπει επίσης την εύκολη προσαρμογή του κατωφλίου συνεισφοράς των καναλιών, δίνοντας έτσι την δυνατότητα να ελεγχθεί το επίπεδο αποσυσχέτισης (disentanglement) του χειρισμού εικόνας.

1.2.3 Stable Diffusion

Το Stable Diffusion [4] είναι ένα μοντέλο λανθάνουσας διάχυσης (Latent Diffusion Model) που επεκτείνει την αρχική αρχιτεκτονική του μοντέλους διάχυσης εισάγοντας έναν αυτοκωδικοποιητή που μεταφέρει τη διαδικασία διάχυσης σε έναν λανθάνοντα χώρο χαμηλότερης διάστασης. Τα νεοεισαχθέντα μοντέλα κωδικοποίησης που αποτελούνται από έναν κωδικοποιητή E και έναν αποκωδικοποιητή D, παρέχουν πρόσβαση σε έναν αποτελεσματικό, χαμηλής διάστασης λανθάνοντα χώρο στον οποίο αφαιρούνται οι υψηλής συχνότητας, ανεπαίσθητες λεπτομέρειες της εικόνας. Αυτή η τροποποίηση παρέχει μια πιο αποδοτική και αποτελεσματική προσέγγιση για τη δημιουργία εικόνων υψηλής ανάλυσης, καθώς μειώνει την πολυπλοκότητα του μοντέλου, διατηρώντας παράλληλα την ποιότητα των παραγόμενων εικόνων.

Για να καταστεί δυνατή η δημιουργία εικόνων υπό συνθήκη, (conditional image generation), το Stable Diffusion κάνει χρήση ενός μηχανισμού διασταυρούμενης προσοχής (cross-attention). Αυτός επιτρέπει στο μοντέλο να εστιάζει επιλεκτικά στις περιοχές της εικόνας που είναι σχετικές με τις πληροφορίες εισόδου, οδηγώντας σε ακριβέστερη και συνεκτικότερη παραγωγή εικόνας. Η προσέγγιση αυτή επιτρέπει τον έλεγχο της παραγόμενης εικόνας από εισόδους διαφορετικών ειδών όπως κείμενο, άλλες εικόνες και σημασιολογικούς χάρτες.

1.3 Μεθοδολογία

Η παρούσα εργασία προτείνει μία νέα μεθοδολογία για την μη επιβλεπόμενη ανακάλυψη ερμηνεύσιμων λανθανουσών κατευθύνσεων στην λανθάνοντα χώρο του Glow, καθοδηγούμενη από φυσική γλώσσα. Η προσέγγισή μας αξιοποιεί τις δυνατότητες δημιουργίας και χειρισμού εικόνων του StyleCLIP για τη δημιουργία ενός συνόλου συνθετικών εικόνων από μια κειμενική περιγραφή. Αυτές οι εικόνες χρησιμοποιούνται στη συνέχεια για τον υπολογισμό της επικρατούσας λανθάνουσας κατεύθυνσης στο Glow, η οποία αντιστοιχεί στο σημασιολογικό χαρακτηριστικό που περιγράφεται από το κείμενο εισόδου. Η προτεινόμενη μέθοδος, με βάση την έρευνα μας, είναι η πρώτη μέθοδος ανακάλυψης λανθανουσών κατευθύνσεων στον χώρο του Glow που καθοδηγείται εξολοκλήρου από μια κειμενική περιγραφή.

Επιπλέον, διερευνούμε τις δυνατότητας σύνθεσης εικόνας από κείμενο του Stable Diffusion για διαφορετικές έννοιες και περιοχές της κατανομής του. Χρησιμοποιούμε την ιεραρχία γνώσης του WordNet για να διερευνήσουμε συστηματικά πώς ιεραρχικά συνδεδεμένες μεταξύ τους έννοιες αναπαρίστανται στο χώρο των εικόνων του Stable Diffusion, και επιχειρούμε να ποσοτικοποιήσουμε αυτές τις ιδιότητες. Η μελέτη μας παρέχει πληροφορίες σχετικά με τις σημασιολογικές ιδιότητες του χώρου εικόνας του Stable Diffusion και μας επιτρέπει να εντοπίσουμε πιθανώς αδύναμες περιοχές της κατανομής.

1.3.1 Glow: Εξερεύνηση Λανθάνοντος Χώρου

Παραγωγή και Τροποίηση Εικόνων

Η προτεινόμενη μεθοδολογία για την ανακάλυψη ερμηνεύσιμων λανθανουσών κατευθύνσεων στο Glow ξεκινά με την παροχή ενός ζεύγους κειμενικών περιγραφών, στο οποίο αναφερόμαστε ως είσοδο. Η είσοδος περιλαμβάνει ένα κείμενο αναφοράς και ένος κειμένο στόχου τα οποία καθορίζουν το επιθυμητό χαρακτηριστικό για το οποίο αναζητούμε την λανθάνουσα κατεύθυνηση στο χώρο του Glow. Η λανθάνουσα κατεύθυνση Δs στο χώρο στυλ του StyleGAN λαμβάνεται μέσω της μεθόδου εύρεσης παγκόσμιων κατευθύνσεων του StyleCLIP. Μετά την απόκτηση του Δs, παράγουμε ένα σύνολο συνθετικών εικόνων D_{source} με χρήση του StyleGAN ενώ παράλληλα αποθηκεύουμε τους αντίστοιχους λανθάνοντες κώδικες του χώρου style space. Στην συνέχεια, προσθέτοντας το διάνυσμα Δs στους λανθάνοντες κώδικες των εικόνων του Δs και παράγουμε τις εικόνες από τους προκύποντες κώδικες. Με τον τρόπο αυτό λαμβάνουμε ένα σύνολο εικόνων D_{target} που αποτελείται εικόνες κατάλληλα τροποποιημένες ώστε να συμμορφώνεται με την κειμενο στόχο της εισόδου.

Η μέθοδος εύρεσης παγκοσμίων κατευθύνσεων του StyleCLIP παραμετροποιείται από δύο μεταβλητές: την ένταση χειρισμού, η οποία ελέγχει την κλίμακα της προκύπτουσας λανθάνουσας κατεύθυνσης, και το κατώφλι αποσυσχέτισης, το οποίο επηρεάζει τον αριθμό των καναλιών του χώρου style space που επηρρεάζονται από τον χειρισμό. Ένα υψηλότερο κατώφλι αποσυσχέτισης οδηγεί σε λιγότερα χειραγωγούμενα κανάλια και υψηλότερη αποδιαπλοκή, γεγονός που σημαίνει ότι άλλα χαρακτηριστικά της εικόνας εκτός του χαρακτηριστικού στόχου μένουν ανεπηρρεάστα.

(a) D_{source} (b) D_{target}

Figure 1.3.1: Δείγματα εικόνων από τα σύνολα D_{source} και D_{target}. Το D_{target} παράχθηκε από την μέθοδο εύρεσης παγκοσμίων κατευθύνσεων με το κείμενο στόχου: "a person with yellow blonde hair"

Οι χειρισμοί που επηρρεάζουν λεπτομερή χαρακτηριστικά της εικόνας γενικά απαιτούν επιλογή μεγάλου κατωφλίου, ενώ για μεγαλύτερης κλίμακας χαρακτηριστικά μικρότερα κατώφλια είναι ευνοϊκότερα. Ως εκ τούτου, δεν είναι δυνατόν να επιλεγεί ένα μοναδικό καλύτερο ζεύγος τιμών παραμέτρων που να λειτουργεί για όλες τις περιπτώσεις. Για την παρούσα εργασία, υιοθετήσαμε μια ευριστική προσέγγιση για τον εντοπισμό των καταλληλότερων τιμών παραμέτρων για κάθε ερώτημα. Ωστόσο, μια βελτιωμένη προσέγγιση θα ήταν ο προσδιορισμός των παραμέτρων που μεγιστοποιούν το Manipulation Disentanglement Score (MDS) του λαμβανόμενου λανθάνοντος κώδικα Δs του StyleCLIP. Το MDS είναι μια μετρική ποιότητας για τις λανθάνουσες κατευθύνσεις, που περιγράφεται στην ενότητα 1.3.1.

Υπολογισμός Λανθάνουσας Κατεύθυνσης

Για τον υπολογισμό της αντίστοιχης λανθάνουσας κατεύθυνση του Glow, είναι απαραίτητο να υπολογιστούν πρώτα οι αναπαραστάσεις των εικόνων στον λανθάνοντα χώρο του Glow πράγμα που καθίσταται εφικτό χάρη

στην αντιστρεψιμότητα του μοντέλου. Αφού ληφθούν οι λανθάνουσες κωδικοποιήσεις $z_{s,1}, z_{s,2}, ..., z_{s,N}$ για το σύνολο δεδομένων D_{source} και $z_{t,1}, z_{t,2}, ..., z_{t,N}$ για το σύνολο δεδομένων D_{target}, η επικρατούσα λανθάνουσα κατεύθυνση Δz μεταξύ των δύο συνόλων μπορεί να υπολογιστεί χρησιμοποιώντας την ακόλουθη έκφραση:

$$\Delta z = \frac{1}{N}\sum_{i=1}^{N} \Delta z_i, \text{ όπου } \Delta z_i = z_{t,i} - z_{s,i} \text{ για } i = 1, 2, \ldots, N$$

Αφού λάβουμε τη λανθάνουσα κατεύθυνση Δz που αντιστοιχεί στην κειμενική περιγραφή στόχου, μπορούμε τώρα να χρησιμοποιήσουμε το Glow για να εκτελέσουμε χειρισμούς διαφορετικής έντασης προσθέτοντας κλιμακωμένες εκδόσεις Δz στον λανθάνοντα κώδικα οποιασδήποτε εικόνας.

Βαθμός Απόσυσχέτισης Λανθάνουσας Κατεύθυνσης

Ο βαθμός αποσυσχέτισης λανθάνουσας κατεύθυνσης (Manipulation Disentanglement Score) ή ΒΑΛΚ είναι μια μετρική αξιολόγησης μιας λανθάνουσας κατεύθυνσης Δz η οποία στηρίζεται στην ποιότητα των μετασχηματισμών εικόνων που προκύπτουν από αυτή. Η μετρική λαμβάνει υπόψιν κατά πόσον ο μετασχηματισμός επιτυγχάνει στην τροποποίηση του επιθυμητού χαρακτηριστικού της εικόνας (ακρίβεια μετασχηματισμού) καθως και σε ποιό βαθμό επηρεάζονται χαρακτηριστικά διαφορετικά του επιθυμητού (αποσυσχέτιση μετασχηματισμού). Για τον υπολογισμό τις μετρικής εφαρμόζουμε μετασχηματισμούς σταδιακά αυξανόμενης έντασης και για κάθε επίπεδο έντασης υπολογίζουμε την ακρίβεια και την αποσυσχέτιση. Ορίζουμε ως ΒΑΛΚ την επιφάνεια που βρίσκεται κάτω από την καμπύλη που σχηματίζουν τα ζεύγη των μετρήσεων ακρίβειας και αποσυσχέτισης όπως απεικονίζεται στο Σχήμα 1.3.2. Για τον υπολογισμό της ακρίβειας και της αποσυσχέτισης, χρησιμοποιούμε προεκπαιδευμένους ταξινομητές 40 διαφορετικών χαρακτηριστικών προσώπου, με σκοπό να εντοπίσουμε την ύπαρξη η όχι των χαρακτηριστικών αυτών στις τροποποιημένες εικόνες.

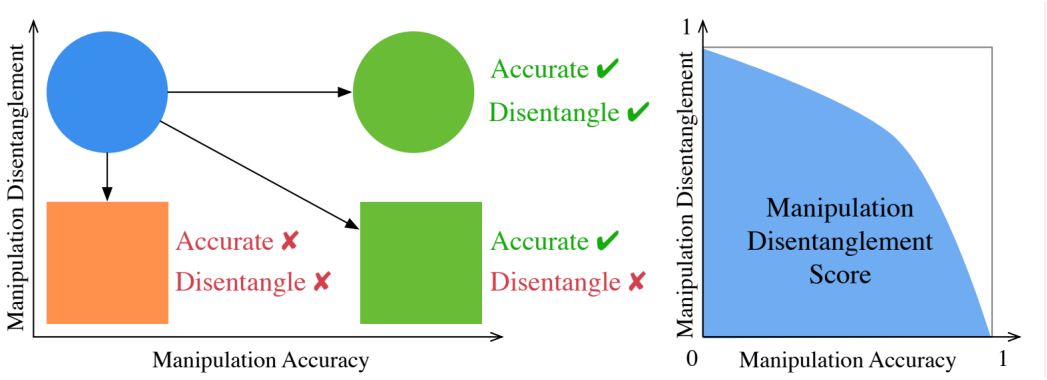


Figure 1.3.2: Οπτικοποίηση του βαθμού αποσυσχέτισης λανθάνουσας κατεύθυνσης [15]

1.3.2 Ανάλυση Stable Diffusion

Η ενότητα αυτή περιγράφει μία μεθοδολογία που επιτρέπει την συστηματική αξιολόγηση της ικανότητας του Stable Diffusion (SD) να παράγει κείμενο από κειμενικές περιγραφές. Πιο συγκεκριμένα, επικεντρωνόμαστε στην ικανότητα του SD να παράγει ακριβή αποτελέσματα για συγγενικές έννοιες και λέξεις. Επιπλέον, χρησιμοποιώντας ομάδες εννοιών με ιεραρχικές σχέσεις, ληφθέντες από το WordNet, εξετάζουμε κατά πόσο η ιεραρχία αυτή διατηρείται και στον χώρο εικόνας του SD.

Για την λήψη μια ιεραρχίας εννοιών από το WordNet, ξεκινάμε από μια γενική έννοια που αντιπροσωπεύεται από ένα synset στο WordNet και επισκεπτόμαστε αναδρομικά τα υπώνυμα της δημιουργώντας έτσι μια ιεραρχία εννοιών. Περιορίζουμε την εξερεύνησή μας σε μέγιστο βάθος 2 και εξετάζουμε έννοιες με συχνότητα εμφάνισης πάνω από ένα κατώφλι για να αποφύγουμε πολύ μεγάλες ιεραρχίες.

Επόμενο βήμα είναι η δημιουργία των κειμενικών περιγραφών που θα τροφοδοτηθούν στο SD για την σύνθεση εικόνων για κάθε έννοια της ιεραρχίας. Η περιγραφές συνιστούνται από το όνομα της έννοιας και την περιγραφή

που παρέχει το WordNet για αυτές. Επιπλέον, διαπιστώσαμε ότι η προσθήκη της λέξης "photograh" στο τέλος κάθε κειμενικής περιγραφής βελτίωσε την ποιότητα των παραγώμενων εικόνων.

Για να καθορίσουμε τις παραμέτρους σύνθεσης εικόνων του SD, πραγματοποιήσαμε μια αναζήτηση πλέγματος. Οι παράμετροι που τροποποιήθηκαν είναι ο αλγόριθμος δειγματοληψίας, τα βήματα δειγματοληψίας και η κλίμακα καθοδήγης του ταξινομητή (CFG). Η κλίμακα CFG ελέγχει το βαθμό συμμόρφωσης της παραγώμενης εικόνας με την κειμενική περιγραφή της εισόδου και επομένως επηρρεάζει την ποικιλομορφία των αποτελεσμάτων. Για να εξαλείψουμε την τυχαιότητα κατά την διάρκεια της αναζήτησης παραμέτρων, θέσαμε ένα σταθερό seed. Οι τελικές τιμές των παραμέτρων που χρησιμοποιήθηκαν για τη δημιουργία εικόνων ήταν οι ακόλουθες:

Parameter	**Value**
Sampler	LMS
Sampling Steps	35
CFG Scale	5.5

Μετά την σύνθεση των εικόνων, χρησιμοποιήσαμε το μοντέλο VGG16 [16] προεκπαιδευμένο στο ImageNet [17] για την αναπαράσταση των εικόνων σε μορφή διανύσματων χαμηλής διαστατικότητας, τα οποία χρησιμοποιούνται για τα πειράματα που ακολουθούν.

Ταξινόμηση σε Υπερώνυμα

Με αυτό το πείραμα, στόχος μας ήταν να ελέγξουμε αν το σημασιολογικό περιεχόμενο των παραγόμενων εικόνων ακολουθεί την ιεραρχία που υπαγορεύεται από τα αντίστοιχες έννοιες στο WordNet. Συγκεκριμένα, ελέγχουμε σε ποιό βαθμό η κατανομή των εικόνων που παράγονται από μια έννοια είναι υπερσύνολο της κατανομής των εικόνων που παράγονται από τα υπώνυμα της έννοιας αυτής.

Για να ποσοτικοποιήσουμε αυτή την ιδιότητα, εκπαιδεύσαμε πρώτα έναν ταξινομητή σε ένα σύνολο δεδομένων εικόνων D_s που δημιουργήθηκε από ένα σύνολο εννοιών T_s που έχουν τουλάχιστον ένα υπώνυμο στην ιεραρχία του WordNet που χρησιμοποιούμε. Ο ταξινομητής που εκπαιδεύτηκε έχει ως είσοδο την διανυσματική αναπαράσταση της εικόνας και ως έξοδο την έννοια από την οποία παράχθηκε η εικόνα. Στη συνέχεια, αξιολογήσαμε τον ίδιο ταξινομητή σε ένα σύνολο δεδομένων εικόνων D_h που δημιουργήθηκε από το σύνολο των υπωνύμων των εννοιών του συνόλου T_s's (T_h). Μια υψηλή ακρίβεια ταξινόμησης θα υποδείκνυε ότι οι κατανομές κάθε έννοιας και των υπωνύμων της αναπαριστάνται με ακρίβεια στο χώρο εικόνων του SD.

Για την εκπαίδευση του ταξινομητή χρησιμοποιήσαμε το auto-sklearn το οποίο είναι ένα πακέτο αυτόματοποιημένης μηχανικής μάθησης. Το εργαλείο αυτό μας επέτρεψε να εξετάσουμε αποδοτικά πολλές διαφορετικές αρχιτεκτονικές για προεπεξεργασία και ταξινόμηση των δεδομένων.

Συσταδιοποίηση Υπωνύμων

Ο στόχος αυτού του πειράματος ήταν να αξιολογηθεί κατά πόσον οι εικόνες που δημιουργούνται από σημασιολογικά παρόμοιες έννοιες παραμένουν κοντά η μία στην άλλη στο χώρο των εικόνων του SD. Για να το πετύχουμε αυτό, πραγματοποιήσαμε μη επιβλεπόμενη συσταδιοποίηση στο σύνολο εικόνων D_h που δημιουργήθηκαν από τα υπώνυμα T_h των εννοιών του T_s. Στη συνέχεια συγκρίναμε τις ομάδοποιήσεις που πήραμε από τον αλγόριθμο με τις πραγματικές ομάδες.

Για να μετρήσουμε την ομοιότητα μεταξύ των προβλεπόμενων και των πραγματικών ομαδοποιήσεων, χρησιμοποιήσαμε τον προσαρμοσμένο δείκτη τυχαιότητας (adjusted rand index), ο οποίος υπολογίζει ένα μέτρο ομοιότητας μεταξύ δύο ομαδοποιήσεων εξετάζοντας όλα τα ζεύγη δειγμάτων και μετρώντας τα ζεύγη που εντάσσονται στις ίδιες ή διαφορετικές ομάδες στην προβλεπόμενη και την πραγματική ομαδοποίηση.

1.4 Αποτελέσματα

1.4.1 Glow

Στην ενότητα αυτή παρουσιάζεται μια σύγκριση μεταξύ της προτεινόμενης μεθόδου έρευσης λανθανουσών κατευθύνσεων του Glow και των καθιερωμένων μεθόδων, με βάση τόσο ποιοτικά όσο και ποσοτικά αποτελέσματα.

Ποιοτητική Αξιολόγηση

Για να έχουμε ένα σημείο αναφοράς για σύγκριση, στην παρούσα μελέτη υλοποιήσαμε τη μέθοδο εύρεσης λανθανουσών κατευθύσεων που χρησιμοποιούν οι δημιουργοί του Glow [8], την οποία αναφέρουμε ως GlowCeleb. Στο GlowCeleb, τα σύνολα εικόνων D_{source} και D_{target} συνοστίζονται από εικόνες του CelebA-HQ, το οποίο είναι ένα σύνολο εικόνων προσώπων με ετικέτες που σηματοδοτούν ύπαρξη διαφόρων χαρακτηριστικών στο πρόσωπο. Η προσέγγιση αυτή εξασφαλίζει ότι τα σύνολα D_{source} και D_{target} αποτελούνται από μεγάλο πλήθος εικόνων με τις επιθυμητες ιδιότητες, αλλά περιορίζει τους πιθανούς χειρισμούς στις ετικέτες χαρακτηριστικών που χρησιμοποιούνται στο CelebA-HQ.

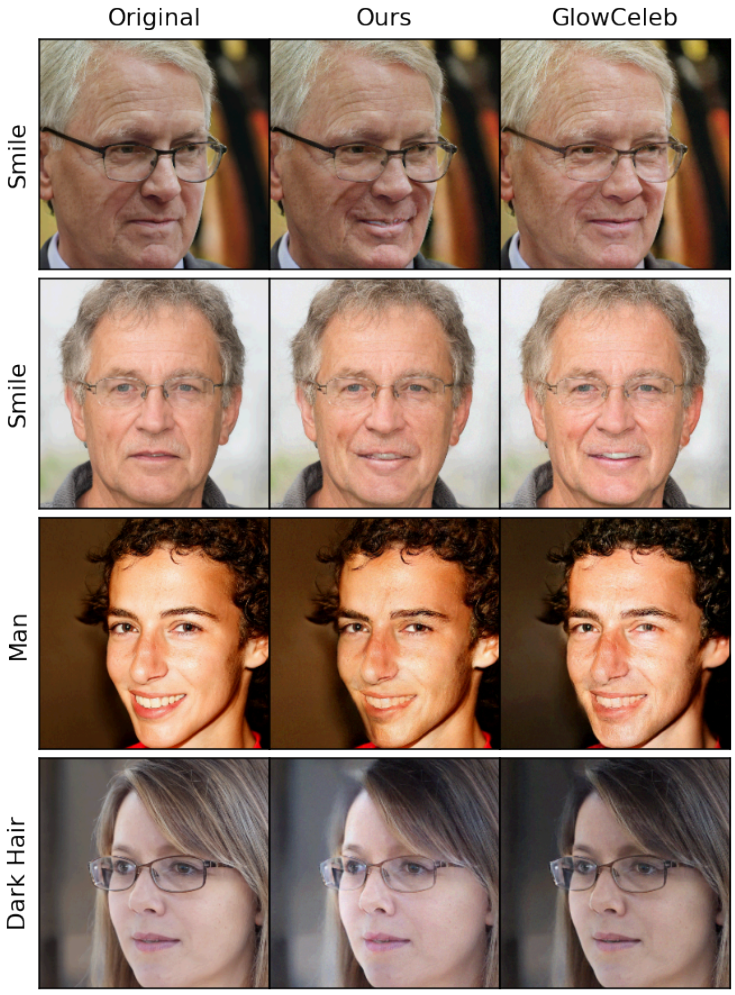


Figure 1.4.1: Σύγκριση της προτεινόμενης μεθόδου εύρεσης λανθανουσών κατευθύνσεων με την μέθοδο GlowCeleb

Παρά τα προαναφερθέντα πλεονεκτήματα του GlowCeleb, όπως η χρήση επισημασμένων δεδομένων για τον προσδιορισμό της αντίστοιχης λανθάνουσας κατεύθυνσης, η μέθοδος μας προσφέρει συγκρίσιμη ποιότητα χειρισμού, ενώ παρέχει πρόσθετη ευελιξία, καθώς η ανακάλυψη της κατεύθυνσης ελέγχεται εξ'ολοκλήρου από την κειμενική περιγραφή της εισόδου. Στο Σχήμα 1.4.1 απεικονίζονται ενδεικτικές τροποποιήσεις εικόνων με τους δύο αυτούς αλγορίθμους οι οποίες υποστηρίζουν τα συμπεράσματα μας.

Ποσοτική Αξιολόγηση

Σε αυτή την ενότητα, παρουσιάζουμε μια ποσοτική αξιολόγηση του μοντέλου μας υπολογίζοντας και συγκρίνοντας το ΒΑΛΚ για διαφορετικά χαρακτηριστικα και διαφορετικές μεθόδους ανακάλυψης λανθάνουσας κατεύθυνσης.

Table 1.1: ΒΑΛΚ με χρήση ταξινομητών χαρακτηριστικών προσώπου. Αναγράφεται και το γεννητικό μοντέλο το οποίο χρησιμοποιήθηκε για τον χειρισμό της εικόνας καθώς και το σύνολο δεδομένων στο οποίο έχει εκπαιδευτεί. Υψηλότερο ΒΑΛΚ είναι καλύτερο.

Μέθοδος	**Μοντέλο**	**Δεδομένα Εκπ.**	**Βαθμός Αποσυσχέτισης ↑**				
			Smile	**Young**	**Dark Hair**	**Gender**	**Overall**
DisCo	StyleGAN2	FFHQ	0.68	0.51	-	-	0.60
GANSpace	StyleGAN2	FFHQ	0.24	-	0.54	0.84	0.54
InterfaceGAN	Progressive-GAN	CelebA-HQ	0.85	-	0.88	0.88	0.87
SGF	Progressive-GAN	CelebA-HQ	0.90	-	0.90	0.88	0.89
GlowCeleb	Glow	CelebA-HQ	0.66	0.88	0.75	0.88	0.79
Ours	Glow	CelebA-HQ	0.60	0.89	0.58	0.95	0.75

Ο πίνακας 8.1 παρουσιάζει το ΒΑΛΚ για τέσσερα χαρακτηριστικά του προσώπου. Τα συγκεκριμένα χαρακτηριστικά επιλέχθηκαν επειδή οι Ren [18] και Li [15] είχαν αναφέρει τα αποτελέσματά τους πάνω σε αυτά. Παρατηρούμε ότι η τιμή του ΒΑΛΚ παρουσιάζει μεγάλη διακύμανση ανάλογα με το εξεταζόμενο χαρακτηριστικό. Όσον αφορά την μέθοδο μας, παρατηρούμε οτι πετυχένουμε σχετικά καλύτερες επιδόσεις σε μεγαλύτερης κλίμακας χαρακτηριστικά του προσώπου, και χαμηλότερες επιδόσεις στα πιο λεπτομερή χαρακτηριστικά. Τελος, σημειώνουμε ότι και στην περίπτωση της ποσοτικής αξιολόγησης η μέθοδος μας πετυχαίνει αποτελέσματα πολύ κοντα με την GlowCeleb.

1.4.2 Stable Diffusion

Πειραματιστήκαμε κυρίως με την ιεραρχία εννοιών που δημιουργήθηκε με ρίζα την έννοια "dog". Αφού παράγαμε τις εικόνες με την μεθοδολογία που περιγράφηκε στην προηγούμενη ενότητα, προχωρήσαμε στην εξαγωγή χαρακτηριστικών με τη χρήση του VGG16 και την δημιουργία διανυσματικών αναπαραστάσεων για κάθε εικόνα.

Ταξινόμηση

Το μοντέλο ταξινόμησης με τις καλύτερες επιδόσεις, όπως προσδιορίστηκε με την βοήθεια του auto-sklearn ήταν ένα σύνολο (ensemble) ταξινομητών που αποτελούνταν από ταξινομητές τύπου Linear Discriminant Analysis, Extra Trees, Passive Aggressive και Random Forest.

Οι μετρικές ταξινόμησης που προέκυψαν παρουσιάζονται στον πίνακα 8.3 και ερμηνεύονται παρακάτω. Επιπλέον, οι προβλέψεις του ταξινομητή απεικονίζονται στο Σχήμα 8.2.3.

Class	**Dataset D_s**			**Dataset D_h**		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Corgi	1.00	0.92	0.96	0.97	0.62	0.75
Cur	0.90	1.00	0.95	0.04	0.14	0.06
Dalmatian	1.00	1.00	1.00	0.95	1.00	0.97
Griffon	0.95	0.88	0.91	0.05	0.18	0.08
Hunting Dog	0.72	0.78	0.75	0.53	0.33	0.41
Poodle	0.97	0.91	0.94	0.99	0.98	0.99
Spitz	0.94	0.94	0.94	0.59	0.99	0.74
Toy Dog	0.85	0.89	0.87	0.75	0.66	0.70
Working Dog	0.71	0.74	0.73	0.80	0.44	0.57
Accuracy	-	-	0.89	-	-	0.57
Macro Avg.	0.89	0.90	0.89	0.63	0.59	0.59
Weighted Avg.	0.90	0.89	0.89	0.71	0.57	0.61

Table 1.2: Μετρικές ταξινόμησης για τα σύνολα D_s και D_h. Σημειώνουμε ότι ο ταξινομητής εκπαιδεύτηκε σε ένα υποσύνολο του D_s.

Βασίσαμε τα συμπεράσματά μας κυρίως στους σταθμισμένους μέσους όρους των μετρικών, δεδομένης της ανισορ-

ροπίας των κλάσεων που υπάρχει στο σύνολο αξιολόγησης D_h. Η ανάλυσή μας δείχνει ότι ο ταξινομητής αποδίδει εξαιρετικά καλά στη μοντελοποίηση της κατανομής του D_s . Ωστόσο, οι βαθμολογίες του ταξινομητή στο D_h, το οποίο αντιπροσωπεύει εικόνες που δημιουργούνται με κάποιο υπώνυμο έννοιας του T_s, δεν είναι το ίδιο υψηλές. Αυτό σημαίνει ότι η κατανομή εικόνων ορισμένων έννοιών της ιεραρχίας, δεν είναι υπερσύνολο της κατανομής εικόνων των υπωνύμων της. Για να ποσοτικοποιήσουμε το βαθμό στον οποίο η κατανομή των εικόνων μιας έννοιας περιέχει την κατανομή των εικόνων των υπονύμων της, μπορούμε να χρησιμοποιήσουμε τη μετρική recall της έννοιας αυτής στο σύνολο ελέγχου D_h.

Ο σταθμισμένος μέσος όρος της μετρικής recall για το σύνολο δοκιμών D_h είναι 0.57 πράγμα που υποδεικνύει ότι στην ιεραχία που εξετάζουμε, η κατανομή εικόνων που προκύπτει από μία έννοια περιλαμβάνει κατά μέσο όρο το 57% της κατανομής εικόνων που προκύπτουν από τα υπώνυμα της έννοιας αυτής. Αυτή η μετρική μπορεί να θεωρηθεί ως μια εκτίμηση του πόσο αποτελεσματικό είναι το Stable Diffusion στην κωδικοποίηση της ιεραρχία του WordNet για αυτή την συγκεκριμένη ομάδα εννοιών.

Συσταδιοποίηση

Για την βελτιστοποίηση των υπερπαραμέτρων του αλγορίθμου συσταδιοποίησης και του σταδίου προεπεξεργασίας των δεδομένων, πραγματοποιήσαμε μια αναζήτηση πλέγματος. Μέσω αυτής της διαδικασίας, αναπτύξαμε ένα μοντέλο που πέτυχε προσαρμοσμένο σκορ rand ίσο με 0.51, υποδεικνύοντας μια μέτρια συσχέτιση μεταξύ των ομάδων εικόνων και των εννοιών που τις παρήγαγαν.

Απεικονίζουμε το δενδρόγραμμα που αντιστοιχεί στη διαδικασία ιεραρχικής ομαδοποίησης στο Σχήμα 1.4.2. Τα ονόματα των φύλλων έχουν τη μορφή έννοια/υπώνυμο. Ωστόσο, η ομαδοποίηση εκτελέσθηκε μόνο με εικόνες υπωνύμωνν (D_h) Συμπεριλαμβάνουμε την έννοια-υπερώνυμο κάθε υπωνύμου για να δείξουμε οπτικά ότι ο αλγόριθμος μπορεί να ομαδοποιήσει επιτυχώς τα υπωνύμα.

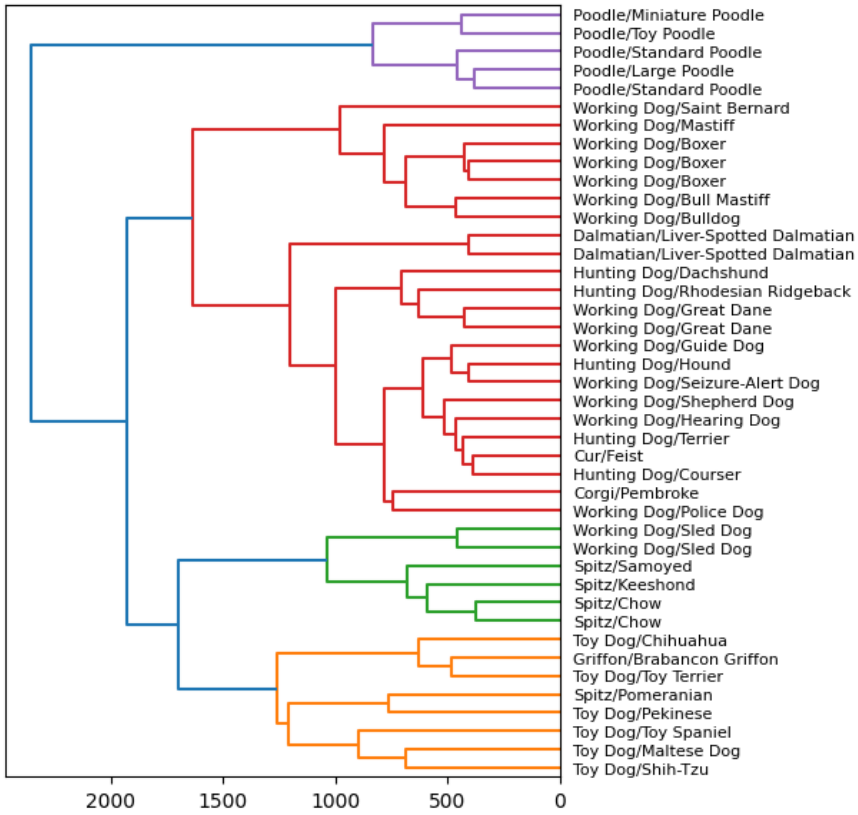


Figure 1.4.2: Δενδρόγραμμα Ιεραρχικής Συσταδιοποίησης

1.5 Συμπεράσματα

1.5.1 Συζήτηση

Στην παρούσα εργασία, προτείναμε μια, νέα μη επιβλεπόμενη μέθοδο για την ανακάλυψη λανθανουσών κατευθύνσεων στον λανθάνοντα χώρο του Glow. Η μέθοδός μας αξιοποιεί μια κειμενική περιγραφή για την ανακάλυψη μιας λανθάνουσας κατεύθυνσης που επηρρεάζει το περιγραφόμενο σημασιολογικό χαρακτηριστικό της εικόνας. Από όσο γνωρίζουμε, δεν υπάρχουν άλλες μέθοδοι εξερεύνησης του λανθάνοντα χώρου του Glow, οι οποίες είναι ελεγχόμενες από φυσική γλώσσα. Παλαιότερες προσεγγίσεις χρησιμοποίησαν μη επιβλεπόμενες ή επιβλεπόμενες μεθόδους μεθόδους για την εύρεση λανθάνουσας κατεύθυνσης. Οι μη επιβλεπόμενες μέθοδοι, συνήθως δεν έχουν τη δυνατότητα αυτόματου ονοματισμού των κατευθύνσεων και απαιτούν ανθρώπινη παρέμβαση για την αντιστοίχιση των σημασιολογικών χαρακτηριστικών σε αυτές. Η δική μας προσέγγιση, καθώς καθοδηγείται από κείμενο, παρακάμπτει αυτά τα ζητήματα ενώ επιτυγχάνει αποτελέσματα συγκρίσιμα με άλλες σύγχρονες μεθόδους ανακάλυψης κατευθύνσεων. Τα ευρήματα αυτά υποστηρίζονται από ποσοτικές και ποιοτικές αξιολογήσεις και συγκρίσεις με άλλες μεθόδους εξερεύνησης του λανθάνοντα χώρου.

Επιπλέον, διερευνήσαμε με συστηματικό τρόπο το μοντέλο Stable Diffusion, ως προς τον τρόπο που μοντελοποιεί και απεικονίζει έννοιες. Χρησιμοποιήσαμε την ιεραρχική γνώση του WordNet για να εξετάσουμε σε ποιό βαθμό το Stable Diffusion μπορεί να απεικονίσει έννοιες που σχετίζονται με μια ιεραρχική σχέση. Τα πειράματά μας έδειξαν ότι για ένα υποσύνολο του WordNet, το Stable Diffusion κατάφερε να διαφοροποιήσει επιτυχώς στενά συνδεδεμένες έννοιες, όπως αποδεικνύεται από τα αποτελέσματα του πειράματος συσταδιοποίησης. Ωστόσο, διαπιστώσαμε ότι η ιεραρχία του WordNet δεν μοντελοποιείται πάντα σωστά από το Stable Diffusion. Για ορισμένα ζεύγη εννοιών E και των υπονύμων τους Y, οι παραγώμενες εικόνες απο την έννοια E δεν ακολουθούν μια κατανομή που είναι κατ'ανάγκη ευρύτερη από την κατανομή των εικόνων των υπονύμων Y. Το γεγονός αυτό, υποδεικνύει ότι το Stable Diffusion, δεχόμενο την έννοια E, έχει μια προκατάληψη υπέρ ορισμένων υπωνύμων και κατά άλλων. Η ιδιότητα αυτή ποσοτικοποιήθηκε για ένα υποσύνολο του WordNet που παράχθηκε με ρίζα την έννοια "dog".

1.5.2 Μελλοντικές Κατευθύνσεις

Στο σημείο αυτό, θα θέλαμε να προτείνουμε ορισμένες πιθανές κατευθύνσεις για την περαιτέρω βελτίωση και εξέλιξη αυτής της εργασίας.

Όσον αφορά τη μέθοδο ανακάλυψης λανθανουσών κατεύθυνσεων, η χρήση ενός διαφορετικού μοντέλου παραγωγής εικόνας απο κείμενο στην θέση του StyleCLIP θα μπορούσε πιθανώς να οδηγήσει σε καλύτερα αποτελέσματα. Νεότερα μοντέλα παραγωγής εικόνας από κείμενο, όπως τα μοντέλα διάχυσης θα αποτελούσαν μια ενδιαφέρουσα εναλλακτική. Επιπλέον, οι κειμενικές περιγραφές που χρησιμοποιήθηκαν για την ανακάλυψη λανθανουσών κατευθύνσεων στην παρούσα μελέτη ήταν σχετικά απλές. Παρόλα, η δυνατότητα καθοδήγησης της μεθόδου μας από κείμενο, επιτρέπει την χρήση πιο περίπλοκών περιγραφών εισόδου για την τροποποίηση σύνθετων χαρακτηστικών της εικόνας. Μια αξιολόγηση της μεθόδου μας σε αυτές τις συνθήκες θα μας επέτρεπε να κατανοήσουμε καλύτερα το εύρος των δυνατοτήτων και τις αδυναμίες της. Ακόμα, θα μπορούσαμε να βελτιώσουμε την προτεινόμενη μέθοδο, επιλέγοντας αυτόματα τις παραμέτρους του αλγορίθμου εύρεσης παγκοσμίων κατευθύνσεων του StyleCLIP. Πιο συγκεκριμένα, μπορούμε να επιλέξουμε τον συνδυασμό παραμέτρων που μεγιστοποιεί τον βαθμό αποσυσχέτισης της προκύπτουσας λανθάνουσας κατεύθυνσης.

Σχετικά με μελλοντικές κατευθύνσεις της μελέτης μας για την αξιολόγηση του Stable Diffusion, θα μπορούσαμε να διερευνήσουμε την αποτελεσματικότητα του μοντέλου σε μεγαλύτερα υποσύνολα του WordNet για να αποκτήσουμε μια πιο ολοκληρωμένη εικόνα της απόδοσής του. Το παραπάνω πείραμα θα μπορούσε να εφαρμοστεί επίσης και σε άλλα μοντέλα διάχυσης, ώστε να γίνει μία σύγκριση μεταξύ τους. Επιπλέον, χρήσιμη θα ήταν η αναζήτηση στρατηγικών ώστε να περιοριστούν οι προκαταλήψεις του Stable Diffusion, οι οποίες μπορούν να ανιχνευθούν με την μεθοδολογία μας. Για παράδειγμα, θα μπορούμε να εξισορροπήσουμε την κατανομή των εννοιών στα δεδομένα εκπαίδευσης ή να χρησιμοποιήσουμε πιο σύνθετες συναρτήσεις κόστους που ενθαρρύνουν την ποικιλομορφία και την δικαιοσύνη (fairness).

Chapter 2

Introduction

Artificial intelligence (AI) has revolutionized the way we interact with technology and has shown remarkable progress in various applications such as natural language processing, computer vision, and robotics. Among these applications, image generation is a rapidly evolving area of research that has seen significant advancements in recent years. Image generation refers to the process of creating realistic images from scratch or modifying existing images using AI algorithms. The use of AI in image generation has resulted in the development of powerful tools for artists, designers, and filmmakers to create high-quality visuals with ease. Moreover, those techniques have found applications in various fields, including medicine, entertainment, and gaming. For instance, AI-generated medical images can help doctors diagnose diseases accurately, while AI-generated images in gaming and entertainment industries can provide a more immersive experience for users.

Advancements in image generation have led to the development of several sophisticated models such as Generative Adversarial Networks (GANs) [1], Variational Autoencoders (VAEs) [2] and diffusion models (DMs) [3]. These models can generate high-quality images that are almost indistinguishable from real images. Additionally, they can be trained on a large dataset of images to learn and mimic the style of a particular artist, era, or genre. More recently, the advent of text-assisted image generative models has enabled the generation and editing of images from natural language descriptions, providing a more intuitive and accessible means of creating visual content [4, 5, 6, 7].

2.1 Motivation

Understanding how generative models represent different concepts is crucial for a number of reasons. Many types of generative models learn to represent complex visual concepts using a low-dimensional latent space, where each dimension corresponds to a different feature of the image. By analyzing the latent space, we can gain insights into how the model represents different concepts and use this knowledge to manipulate or generate images with specific attributes. We can also gain insights regarding the limitations and potential biases of generative models. For example, a model trained on a specific dataset may not generalize well to new data or concepts that are not well represented in the training data. Therefore, it is essential to analyze the model's performance on different tasks and evaluate its strengths and weaknesses to determine where it may need improvement.

2.2 Contribution

In this work we utilize text-conditional image generation models such as StyleCLIP [12] and Stable Diffusion [4] to achieve the following:

- Propose a novel method of discovering interpretable latent directions in the feature space of generative models such as Glow [8] using a natural language description.

- Evaluate and compare the discovered latent directions with other state of the art methods, in terms of the quality of the resulting image manipulation.
- Analyze Stable Diffusion's ability to differentiate between closely related textual concepts with the help of WordNet [19].

2.3 Structure

The thesis consists of nine chapters, the first being this introduction.

Chapter 3 discusses generative models, including Generative Adversarial Networks (GANs), Flow-based Generative Models, and Diffusion Models. The chapter provides an overview of these models and their training process, with specific focus on StyleGAN, Glow, and diffusion models.

Chapter 4 focuses on text-guided image generation and covers two state-of-the-art models: StyleCLIP and Stable Diffusion. The chapter discusses how these models generate images based on textual input.

Chapter 5 examines the interpretability of generative models, focusing on the latent space and its exploration. The chapter discusses unsupervised and supervised methods for exploring the latent space.

Chapter 6 provides an overview of WordNet, a lexical database that organizes words based on their semantic relationships.

Chapter 7 describes the methodology used in the study, including the use of Glow and Stable Diffusion for latent space exploration and image generation. The chapter also discusses how the manipulation disentanglement score is computed and how images generated with Stable Diffusion are used for classification and clustering.

Chapter 8 presents the results of the study, including qualitative and quantitative evaluations of the proposed Glow and Stable Diffusion methodologies. The chapter also discusses the classification results obtained with Stable Diffusion.

Finally, Chapter 9 provides the conclusion of the study and highlights the future directions of the research.

Chapter 3

Generative Models

Generative models [20] are a type of statistical model that aims to learn the underlying probability distribution of a given dataset. Given a training set of n observations, represented as $x = x_1, x_2, ..., x_n$, a generative model estimates the joint probability distribution $p(x, \theta)$ of the data x and the model parameters θ. In other words, a generative model learns to model the underlying probability density function of the data.

More formally, let x be a random variable representing the data and θ be a vector of parameters that govern the distribution of the data. A generative model estimates the joint distribution $p(x, \theta)$ such that the likelihood of observing a given data point x_i is given by:

$$p(x_i|\theta) = \int p(x_i|z, \theta)p(z|\theta)dz$$

where $p(x_i|z, \theta)$ is the conditional probability of observing x_i given the latent variable z and the model parameters θ, and $p(z|\theta)$ is the prior probability distribution of the latent variable z.

The objective of a generative model is to estimate the model parameters θ that maximize the likelihood of the observed data:

$$\theta^* = \arg\max_{\theta} p(x|\theta)$$

Once trained, a generative model can be used to generate new samples of the data by sampling from the learned distribution $p(x, \theta)$. These generated samples can be used for a variety of tasks, including data augmentation, data visualization, and data synthesis.

There are several types of generative models, including parametric models such as Gaussian mixture models [21] and non-parametric models such as kernel density estimation [22]. Deep generative models such as Variational Autoencoders (VAEs) [2], Generative Adversarial Networks (GANs) [1], and Latent Diffusion Models (LDMs) [3] are currently the state-of-the-art in generative modeling, thanks to their ability to model complex, high-dimensional data distributions. A high level overview of their architecture is presented in Figure 3.0.1.

Contents

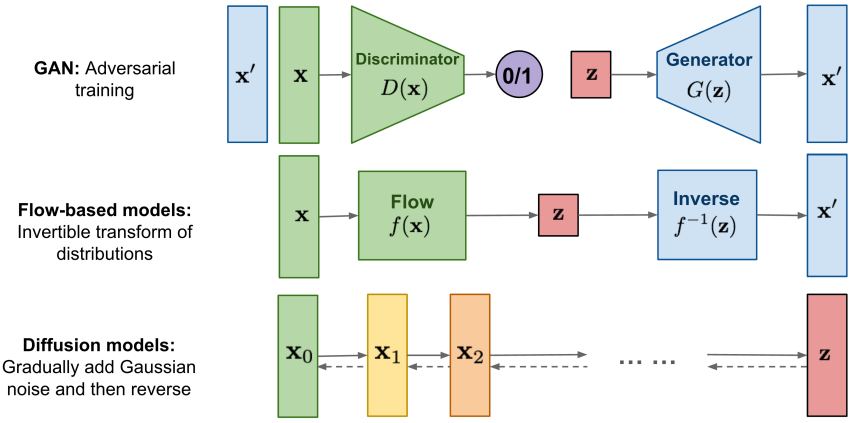


Figure 3.0.1: Overview of the three types of generative models that are used in this work.

3.1 Generative Adversarial Networks

Generative Adversarial Networks, or GANs for short, are a type of generative model in deep learning that have gained a lot of attention and popularity in recent years. GANs were first introduced by Ian Goodfellow in 2014 [1], and have since been used to generate realistic images, videos, and even music. The basic idea behind GANs is to train two neural networks, a generator and a discriminator, in a competitive setting where the generator tries to produce samples that are indistinguishable from the real data, while the discriminator tries to distinguish between the real and fake data. Through this adversarial training process, the generator learns to produce increasingly realistic samples, while the discriminator becomes better at detecting fake samples produced by the generator.

3.1.1 Training

Training a GAN can be thought of as a minimax optimization problem, where the generator aims to minimize the following objective function:

$$min_{\theta_G} max_{\theta_D} V(D,G) = \mathbb{E}_{x \sim p_{data}(x)}[logD(x)] + \mathbb{E}_{z \sim p_z(z)}[log(1 - D(G(z)))]$$

where θ_G and θ_D are the parameters of the generator and discriminator, respectively, $p_{data}(x)$ is the distribution of the real data, $p_z(z)$ is the prior distribution of the noise vector z, and $D(x)$ and $D(G(z))$ are the discriminator outputs for the real data and generated samples, respectively.

The generator is trained to maximize the second term of the objective function, which corresponds to the probability that the discriminator outputs a high value for a generated sample. On the other hand, the discriminator is trained to maximize the first term, which corresponds to the probability that it correctly identifies a real sample as real and a generated sample as fake. The generator and discriminator are trained iteratively via backpropagation, where the discriminator is first trained on a batch of real and fake samples, and then the generator is trained on a new batch of noise vectors.

The training of GANs is known to be challenging, as the generator and discriminator can easily fall into a state of equilibrium where the generator produces low-quality samples that the discriminator cannot distinguish from the real data. This problem is known as mode collapse and can be addressed by using various techniques such as adding noise to the discriminator input or using different loss functions. Despite these challenges, GANs have shown remarkable success in generating realistic images, videos, and even music, and continue to be an active area of research in deep learning.

3.1.2 StyleGAN

StyleGAN is a type of generative model that was introduced in 2018 by Karras et al [14]. It is an extension of the original GANs architecture that is designed to generate high-quality, high-resolution images with improved

control over various aspects of image synthesis. The key innovation of StyleGAN is the incorporation of style-based generator architecture, which allows for greater control over the image synthesis process.

The style-based generator architecture of StyleGAN consists of two major components: a mapping network and a synthesis network. The mapping network takes as input a random noise vector and maps it to a learned style vector that controls various aspects of image synthesis. The synthesis network takes the style vector as input and generates an image.

One of the main advantages of StyleGAN over other generative models is its ability to control various aspects of image synthesis. The style vector output by the mapping network is divided into several different components, each of which controls a different aspect of image synthesis. These components include global styles, which control the overall appearance of the image, and local styles, which control the appearance of specific regions of the image. By manipulating the style vector components, users can control various aspects of image synthesis, including the pose, expression, and identity of the generated images.

Another key feature of StyleGAN is its progressive growing technique, which allows for the generation of high-resolution images. The model is trained on low-resolution images first and then gradually increases the resolution of the generated images as the training progresses. This technique ensures that the generator learns to capture the details of the image at different scales, resulting in high-quality, realistic images.

StyleGAN has been used in a variety of applications, including image editing, fashion, and entertainment. It has also been used to generate high-quality images of faces and objects, with applications in fields such as computer vision, graphics, and art. The model has been shown to generate images that are visually stunning and highly realistic, with a level of control over the synthesis process that was not previously possible with other generative models.

In conclusion, StyleGAN is a powerful generative model that offers improved control over various aspects of image synthesis, including pose, expression, and identity. Its style-based generator architecture and progressive growing technique enable the generation of high-quality, high-resolution images with a level of control that was previously not possible. The model has shown great promise in a variety of applications and continues to be an active area of research in the field of deep learning.

3.2 Flow-based Generative Models

Implicit generative models, such as GANs, do not explicitly model the likelihood function or provide a means for identifying the latent variable corresponding to a particular sample. In contrast, flow-based generative models explicitly model the likelihood function by utilizing normalizing flow, a statistical technique that utilizes the change-of-variable law of probabilities to transform a simple distribution into a complex one. The direct modeling of likelihood offers numerous benefits, including the ability to compute and minimize the negative log-likelihood as the loss function. Moreover, latent variables can inferred and new samples can be generated by sampling from the initial distribution and applying the flow transformation.

3.2.1 Normalizing Flows

Normalizing Flow (NF) [9] models are a powerful approach for distribution approximation. These models transform a simple distribution $p_0(z_0)$ into a complex one $p_n(z_n)$ through a sequence of invertible transformation functions f_i. By flowing through a chain of transformations, the variable z is repeatedly substituted for the new one according to the change of variables theorem, eventually resulting in a probability distribution of the final target variable. Specifically, let $z_1, z_2, \ldots, z_n$ be random variables with respective distributions p_i, and let $z_i = f_i(z_{i-1})$, where f_i are invertible functions. Then the distribution $p(x) = p_n(z_n)$ at the end of the chain is given by the formula:

$$\log p(x) = \log p_0(z_0) - \sum_{i=1}^{n} \log \left| \frac{df_i}{dz_{i-1}} \right|$$

With the assumption that the transformations f_i are easily invertible and their Jacobian determinants can be computed fast, the exact log-likelihood of the data x becomes computationally tractable. Therefore, the

training criterion of a flow-based generative model is the negative log-likelihood over the training dataset D, expressed as:

$$L(D) = -\frac{1}{|D|} \sum_{x \in D} \log p(x)$$

3.2.2 Glow

The Glow model [8] is a deep generative model that implements a normalizing flow by stacking a sequence of invertible bijective transformation functions, building upon the NICE [10] and RealNVP [11] flow models. Specifically, each step of the flow in the Glow model comprises three distinct layers.

The first layer is an activation normalization (actnorm) layer, which normalizes the activation output of the previous layer by scaling and shifting it to have zero mean and unit variance along each channel.

The second layer is an invertible 1 x 1 convolution layer that mixes the channels in the input feature map to allow for interactions between them. This layer has a weight matrix that is initialized as a random orthogonal matrix to ensure that it is invertible.

The third layer is a coupling layer, which partitions the input channels into two groups and applies an invertible transformation to one group based on the other group. This layer ensures that the output remains invertible and preserves the dimensions of the input, while introducing non-linearity in a local fashion.

Those layers are combined in multi-scale architecture as shown in Figure 3.2.1

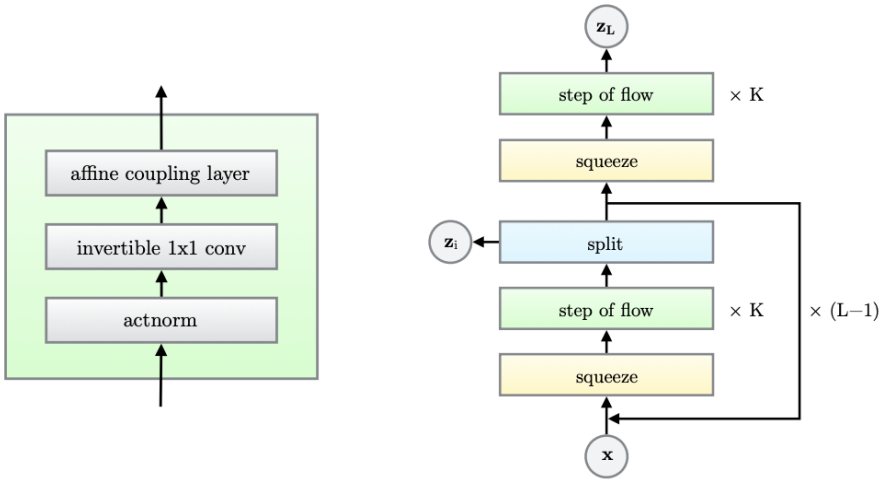


Figure 3.2.1: One step of flow (left) and multiple steps of flow combined into a multi-scale architecture (right) [8]

Glow, like all flow-based models has the ability to perform accurate inference of the latent variables of a given image. This capability enables the transformation from the image space to the latent space, which allows for accurate image manipulations and reconstructions. By manipulating the latent variables, one can precisely control the appearance of the resulting image. This makes Glow model a powerful tool for various image generation and manipulation tasks.

3.3 Diffusion Models

Diffusion models [3] are a type of generative model used for image generation that apply a sequence of denoising transformations to a noise-initialized image to generate a high-quality image. The diffusion process can be represented as a Markov chain, where each state of the chain is an image at a certain stage of the diffusion process, and the final state of the Markov chain is the generated image. More specifically, diffusion models can be interpreted as a sequence of denoising autoencoders $\varepsilon_\theta(x_t, t); t = 1...T$, which are trained to

predict a denoised variant of their input x_t, where x_t is a noisy version of the input x. Their optimization objective can be be expressed as:

$$L(DM) = \mathbb{E}_{x,\varepsilon\sim N(0,1),t}\|\varepsilon - \varepsilon_\theta(x_t, t)\|_2^2$$

with t uniformly sampled from $1, ..., T$.

Diffusion models have the advantage of generating high-quality images with rich textures and details, even for highly complex image datasets. They also allow for fine-grained control of the image synthesis process, enabling the generation of images that satisfy specific constraints, such as image resolution or image style. However, diffusion models rely on a long Markov chain of diffusion steps to generate samples, which can be computationally expensive in terms of time and compute resources. Despite recent developments in the field of diffusion-based image synthesis, such as Stable Diffusion, DALL-E2, Imagen, and DreamBooth the sampling is still slower than other generative modeling methods such as GANs.

Overall, diffusion models represent a promising approach for image generation with several advantages over other methods. As the field of generative modeling continues to advance, diffusion models are likely to play an increasingly important role in generating high-quality images for a range of applications such as unconditional image synthesis, inpainting, super-resolution and text-guided image synthesis.

Chapter 4

Text-Guided Image Generation

Text-conditional image generation is a type of generative modeling in which an algorithm is trained to generate images from textual descriptions. The goal is to create images that are consistent with the given textual description, while also being visually appealing and realistic.

Contents

4.1 StyleCLIP

StyleCLIP [12] uses a combination of language and image representations to guide the image generation process. Specifically, StyleCLIP combines the power of Contrastive Language-Image Pre-training (CLIP) models [13] with StyleGAN 2 [14], to generate images that are both semantically and visually consistent with the given textual description.

The StyleCLIP method works by manipulating the latent space of a pre-trained generative model (StyleGAN2 in this case) using a combination of a textual description and a reference image. The textual description is used to guide the semantic content of the generated image, while the reference image is used to provide additional style information. The objective of this process os to find a latent code that produces an image that is both semantically and visually similar to the given textual description and reference image, respectively.

One of the key advantages of StyleCLIP is its ability to generate images that are highly specific to the given textual description. This is achieved through the use of the CLIP, which allows for more nuanced and contextualized understanding of the textual input. CLIP is a pretrained model that was trained on pairs of images and textual descriptions. It allows for the representation and comparison of images and text into a multi-modal joint embedding space.

4.1.1 Image Manipulation

In the StyleCLIP paper, the authors describe three different methods for manipulating images using text inputs. These methods differ in their approach of incorporating textual information and their level of interpretability.

The first method is the optimization-based approach, which involves finding a latent code that produces an image that is consistent with both the textual input and a reference image. This approach is highly flexible, as it allows for the generation of highly specific images that are consistent with the given input. However, the optimization process can be time-consuming, and the resulting image features may be entangled, making it difficult to adjust specific aspects of the generated image.

The second method is the local direction mapper, which involves training a model to infer a manipulation step in $W+$ from an image's latent representation in $W+$ space. This mapping model is trained for a specific text prompt and infers a manipulation direction that is custom-tailored to the input image.

The third method is the global direction model, which involves learning a mapping from CLIP space to style space. Specifically, this approach uses a pair of source and target text prompts to determine the relevance of each channel of the style space to the desired manipulation direction. The method starts by computing the CLIP-direction that corresponds to the difference between the neutral text prompt and the target text prompt. Next, the CLIP-direction is compared to the direction corresponding to perturbations on each channel of the style space using cosine similarity. Channels with relevance smaller than a selected threshold are ignored, and only highly relevant channels are used to determine the direction Δs in the style space. This direction can then be used to manipulate the generated image consistently with the target text prompt. This approach allows for rapid manipulation of the generated image based on a pair of source and target text prompts, without the need for optimization on a per-image basis or training on a per-text prompt basis. Furthermore, the global direction model also permits easy adjustment of the channel relevance threshold parameter to disentangle features, making it more interpretable than the other two approaches.

Table 4.1: Preprocessing, training and inference times of the three text-driven manipulation methods [14]

Method	Preprocessing Time	Train Time	Inference Time
Optimizer-based	-	-	98 sec
Local Direction Model	-	10-12h	75 ms
Global Direction Model	4h	-	72 ms

After consulting the execution times of Table 4.1 and experimenting with the 3 methods, it was determined that the use of the global direction model was the optimal choice for this study, due to its favorable combination of low computation time, good quality of image manipulations and the flexibility offered around the control of the manipulation's disentnaglement.

4.2 Stable Diffusion

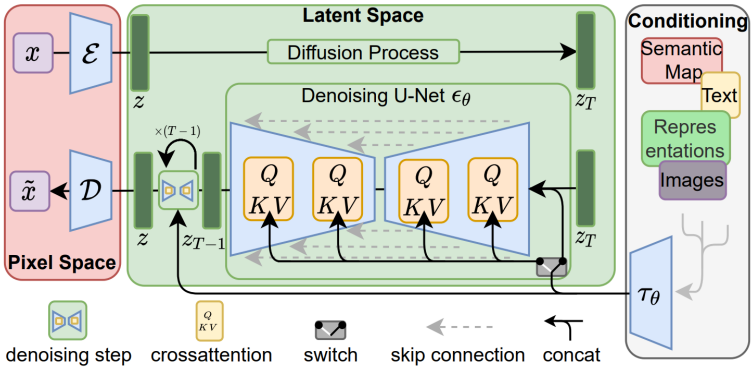


Figure 4.2.1: Stable Diffusion's architecture with encoder, decoder and cross-attention modules

Stable diffusion [4] is a latent diffusion model (LDM) that extends the original DM architecture by introducing an autoencoder that transfers the diffusion process to a lower-dimensional latent space. The newly introduced compression models consisting of an encoder E and a decoder D, give access to an efficient, low-dimensional latent space in which high-frequency, imperceptible details are abstracted away. This alteration provides a more efficient and effective approach to high-resolution image generation, as it reduces the complexity of the model while maintaining the quality of generated images. In this case the objective function is expressed by:

$$L(DM) = \mathbb{E}_{\mathcal{E}(x),\varepsilon\sim N(0,1),t}\|\varepsilon - \varepsilon_\theta(z_t, t)\|_2^2$$

To enable conditional image generation, Stable Diffusion employees a cross-attention mechanism, as depicted in Figure 4.2.1, that allows the model to attend to specific regions of the input information when generating the image. The cross-attention mechanism involves using the input information to compute attention maps, which are then used to weight the features extracted from the image. This allows the model to selectively focus on the regions of the image that are relevant to the input information, leading to more accurate and coherent image generation. This approach is especially effective when conditioning the image generation process with many different modalities such as text, other images or semantic maps.

Figure 4.2.2: Generated image for the text prompt "Master Yoda riding a white horse on Mars"

Chapter 5

Interpretability of Generative Models

5.1 Latent Space

The latent space of generative networks is a fundamental concept in deep learning that has significant implications for image manipulation and synthesis. The latent space refers to the high-dimensional space of latent variables that a generative network uses to map from a prior distribution to a data distribution of interest. These latent variables control various aspects of the generated image, including color, texture, and shape.

One of the key advantages of working in the latent space is the ability to perform targeted image manipulations. By manipulating the latent variables, it is possible to generate new images that differ from the original in specific ways. For instance, by modifying the relevant latent variables, one can alter the pose or expression of a face in an image or change the texture or color of an object. Additionally, interpolating between the latent codes of two images, as shown in Figure x, reveals the ability of the latent space to encode semantically meaningful image information. Image manipulations using the latent space of generative networks have broad applications, including image editing, data augmentation, and data synthesis. In computer vision, generative networks are often used to augment training datasets by generating new images that are similar to the original data but have specific modifications, thereby improving the robustness and generalization of deep learning models.

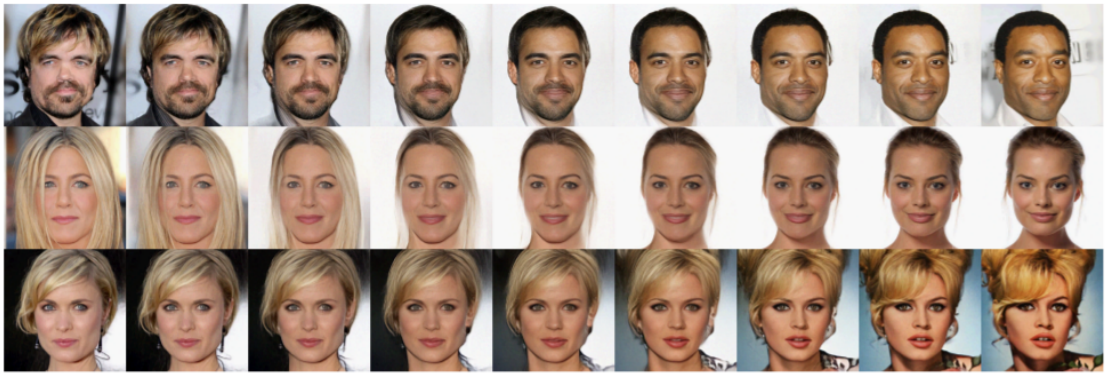


Figure 5.1.1: Linear interpolation in Glow's latent space between real images [8]

Another important concept in the latent space of generative networks is disentanglement. Disentanglement refers to the ability to separate the latent variables into independent and interpretable factors of variation. In other words, disentangled representations of the latent space allow for greater control over specific aspects of the generated image. For example, a disentangled representation of the latent space for faces might separate the pose and expression of the face into different factors of variation. This separation would allow for greater control over the manipulation of these specific attributes.

5.2 Latent Space Exploration

5.2.1 Unsupervised Methods

GANSpace [23] is a straightforward approach, which provides a framework for unsupervised discovery of interpretable controls for GANs. GANSpace computes the principal directions in the W space of StyleGAN by sampling a large number of random noise vectors z and performing principal component analysis (PCA) on the corresponding intermediate latent vectors w. By altering w along the principal directions, GANSpace allows for the exploration of entanglements between different concepts in the generated images. Notably, entanglement occurs when a single direction in the latent space affects multiple semantic aspects of the generated image. To mitigate this issue, GANSpace proposes layer-wise edits, where only certain layers of the generator network are modified, leading to a more disentangled representation of the generated images.

Voynov [24] proposed an optimization-based method for discovering interpretable directions in the GAN latent space. The methods aims to find a set of directions A that lead to more distinguishable transformations in the generated images. This is achieved by jointly optimizing for the directions A and a reconstructor R that aims to determine the transformation direction given an image and its transformed version.

Disentanglement via Contrast, or **DisCo** [18], first generates a large number of synthetic images using a pretrained generative model and then learns a feature extractor using contrastive learning. The feature extractor is trained to map the synthetic images to a latent space where each dimension corresponds to a disentangled factor. The contrastive learning objective encourages the feature extractor to produce similar embeddings for images that share the same disentangled factor and dissimilar embeddings for images that differ in that factor. By exploiting the pretrained generative model, the method can learn disentangled representations without requiring additional annotated data or modifying the architecture of the generative model.

5.2.2 Supervised Methods

InterFaceGAN [25] is a method for interpreting the disentangled face representation learned by GANs. The method assumes that hyperplanes in the latent space are the separation boundary for the existence/absence of semantic features. By adding a vector proportional to the normal vector of the hyperplane corresponding to a specific feature to the noise vector z, the feature can be controlled while keeping other features unaffected. To learn the hyperplanes of a pre-trained GAN, InterFaceGAN uses a pre-trained classifier on 500k synthesized images to obtain attribute scores. Images with high certainty of the presence or absence of attributes are selected for training. A linear SVM is trained to predict attribute scores given the noise vector of the synthesized image, and the resulting hyperplanes are the desired directions. InterFaceGAN measures the entanglement between semantic attributes with corresponding hyperplanes using the dot product of the normal vectors. Attribute correlation is measured by the correlation of attribute scores predicted by the SVM, which is similar between a large number of synthesized images and the training set. For StyleGAN, W space is used instead of Z space.

Yang [26] proposed a method for discovering interpretable directions in the latent space of GANs for attributes that are not binary. Unlike previous methods, it does not assume linear separability and only requires positive samples of the attribute to learn the corresponding direction in the latent space. The method uses adversarial optimization to find the direction ϑ in the latent space that corresponds to the chosen attribute. The optimization is performed with an attribute assessor, discriminator, and identity loss to ensure that the transformed image is realistic and similar to the original. The discovered directions for different attributes show little correlation, allowing multiple attributes to be altered by adding their corresponding ϑs.

Chapter 6

WordNet

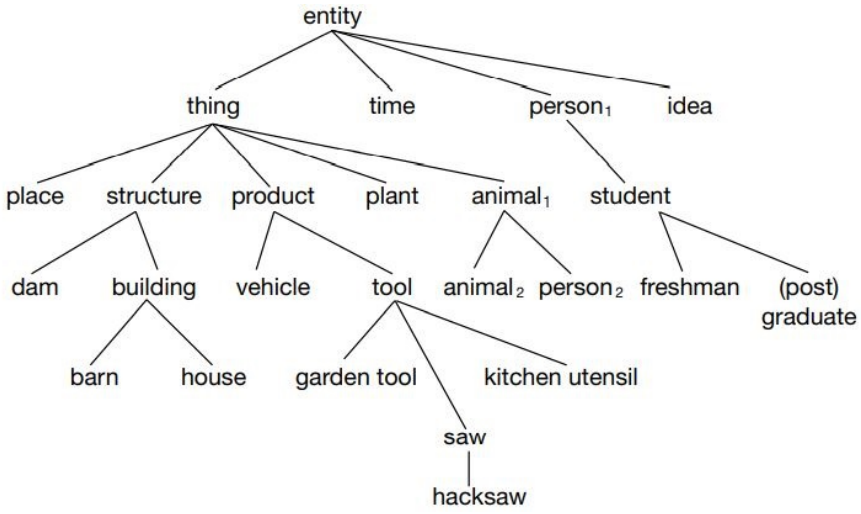


Figure 6.0.1: Hyponym hierarchy for a subset of WordNet

WordNet [19] is a widely-used lexical database that provides a comprehensive understanding of the English language through the grouping of nouns, verbs, adjectives, and adverbs into sets of cognitive synonyms or synsets. Each synset represents a distinct concept, and they are interlinked by conceptual-semantic and lexical relations, creating a network of related words and concepts such as the one depicted on Figure 6.0.1. WordNet's network structure allows it to be used in various fields, including computational linguistics and natural language processing, as it provides a rich resource for semantic analysis.

In WordNet, words are related through various conceptual relations, with the most common relation being synonymy. Synonyms, which are words that denote the same concept and can be used interchangeably in many contexts, are grouped into sets of synsets, each containing a brief definition and one or more short sentences illustrating the use of the synset members. Each form-meaning pair in WordNet is unique, and word forms with multiple meanings are represented by distinct synsets.

The super-subordinate relation, also known as hyperonymy, hyponymy, or ISA relation, is the most frequently encoded relation among synsets. It links more general synsets, such as "furniture" and "piece of furniture," to increasingly specific ones, such as "bed" and "bunkbed." WordNet distinguishes between Types, which are common nouns, and Instances, which are specific persons, countries, and geographic entities. Instances are always leaf (terminal) nodes in their hierarchies. The meronymy relation, which is the part-whole relation, holds between synsets such as "chair" and "backrest" or "seat" and "leg."

Chapter 7

Methodology

This thesis presents a novel framework for unsupervised discovery of interpretable latent directions in the latent space of Glow, guided by natural language. Our approach leverages the image generation and manipulation capabilities of StyleCLIP, a text-conditional model, to create a dataset comprising of manipulated and non-manipulated versions of images. These images are then used to compute the prevalent latent direction in Glow, corresponding to a semantic attribute of an image expressed as a textual description. To the best of our knowledge, this is the first method for discovering latent directions in Glow that can be guided by natural language.

Furthermore, we investigate the variability of Stable Diffusion's text-conditional image generation capabilities across different areas of its sample distribution. We utilize WordNet's hierarchical knowledge to explore how closely related textual concepts and their differences are represented in the image space, and we attempt to quantify these qualities. Our study provides insights into the semantic properties of Stable Diffusion's image space and sheds light on the extent to which textual concepts are reflected in the generated images.

7.1 Glow: Latent Space Exploration

7.1.1 Image Manipulation with StyleCLIP

The proposed methodology for discovering interpretable latent directions in Glow is initiated by providing a pair of textual descriptions, which we refer to as a query. The query comprises a target prompt that specifies the desired attribute and a source prompt that serves as a reference description and should be neutral. The corresponding latent direction Δs in StyleGAN's style space [27] is obtained by feeding the query to StyleCLIP's Global Directions algorithm.

After obtaining Δs, a set of images D_{source} is generated with unconditional image generation using StyleGAN, while storing the corresponding style space codes in the process. By adding Δs to each latent code and generating the corresponding images, a set of images D_{target} is obtained, manipulated to conform to the target prompt.

The Global Directions algorithm is parameterized by two variables: the manipulation strength, which controls the scale of the resulting latent direction, and the disentanglement threshold, which affects the number of style space channels that are manipulated. A higher disentanglement threshold results in fewer manipulated channels and higher disentanglement, implying that other attributes other than the target are not affected as much. Large scale manipulations require lower disentanglement thresholds, whereas manipulations that affect small details require higher thresholds. Hence, it is not possible to select a single best pair of parameter values that work for all cases. For this work, we adopted a heuristic approach to identify the most appropriate parameter values for each query. However, an improved approach would be to determine the optimal parameters that maximize the Manipulation Disentanglement Score of the obtained latent code Δs of StyleCLIP. The Manipulation Disentanglement Score is a quality metric for latent directions, described in section 7.1.3.

(a) D_{source}

(b) D_{target}

Figure 7.1.1: Example images taken from the datasets D_{source} and D_{target}. D_{target} was generated via the Global Direction algorithm with the target prompt: "a person with yellow blonde hair"

It is crucial to ensure that the generated image datasets D_{source} and D_{target} conform to Glow's modeled distribution. In this work, we employed a StyleGAN and Glow model trained on human face datasets, specifically Flickr-Faces-HQ (FFHQ) and CelebA-HQ, respectively. Therefore, even with StyleCLIP's unconditional image generation, the resulting dataset D_{source} falls within the distribution modeled by our Glow model. However, if we were to use a different model, such as Stable Diffusion, the resulting image dataset D_{source} might not fall within the distribution of the much smaller Glow model. This is because Stable Diffusion is trained on a broader and more diverse set of images. To ensure that the generated image dataset D_{source} conforms to the distribution modeled by Glow, we should employ text-conditional image generation when using Stable Diffusion. This ensures that the generated images are more likely to fall within the modeled distribution of Glow, despite the differences in the training data of the two models.

7.1.2 Latent Direction Computation

To obtain the corresponding latent direction in the latent space of Glow, it is necessary to first obtain the representations of the images in the latent space of Glow. This is made possible by the invertibility of Glow, which enables latent code inference. Once the latent codes $z_{s,1}, z_{s,2}, ..., z_{s,N}$ for the D_{source} dataset and $z_{t,1}, z_{t,2}, ..., z_{t,N}$ for the D_{target} dataset have been obtained, the prevalent latent direction Δz between the two datasets can be computed using the following expression:

$$\Delta z = \frac{1}{N}\sum_{i=1}^{N} \Delta z_i, \text{ where } \Delta z_i = z_{t,i} - z_{s,i} \text{ for } i = 1, 2, \ldots, N$$

After obtaining the latent direction Δz which corresponds to the target prompt we can now use Glow to perform manipulations of varying intensity by adding scaled versions of Δz to an images latent code.

7.1.3 Manipulation Disentanglement Score

The manipulation disentanglement score is a metric that is used to evaluate the quality of a latent direction Δz. The score is obtained by adding Δz to the latent code of a large number of images at progressively larger scales, thereby generating manipulated versions of the images with varying manipulation intensities. For each manipulation intensity, CelebA-HQ 40 pre-trained facial attribute classifiers are used on all images to help compute two quantities: the accuracy of the manipulation and the disentanglement measure of the manipulation.

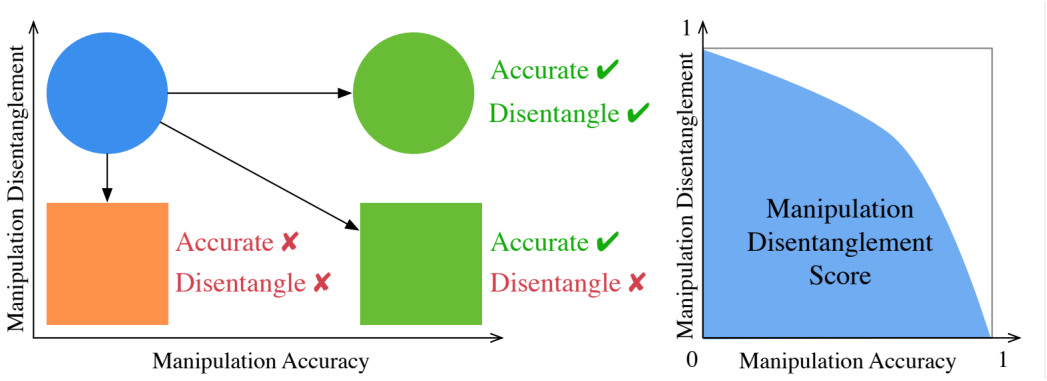


Figure 7.1.2: Illustration of the Manipulation Disentanglement Score [15]

At the ith intensity level, which corresponds to manipulating images by adding $l_i\Delta z$ to their latent codes, the accuracy of the manipulation is defined as the percentage of manipulated images where the target attribute score, as computed by the pre-trained classifiers, increased by a predetermined threshold. The disentanglement of the manipulation is defined as the percentage of attributes, other than the target attribute, that remained unaffected by the manipulation. We compute the accuracies and disentanglement scores at

each intensity level and then plot the disentanglement scores over the accuracies to obtain the Manipulation Disentanglement Curve (MDC). The area contained under this curve is defined as the Manipulation Disentanglement Score (MDS).

In this study, we also adopted a modified version of MDS for evaluation, as initially proposed by Li [15]. Unlike the conventional MDS approach, which employs pre-trained attribute classifiers for disentanglement computation, our method leverages image similarity between the manipulated and non-manipulated images as a reliable proxy. Specifically, we utilized Inception-Resnet-v1 [28] pre-trained on VGGFace2 [29] to extract feature embeddings, followed by calculating the cosine similarity of the respective feature embeddings to get the image similarity.

7.2 Stable Diffusion Analysis

The goal of this section is to systematically explore the text conditional image generation capabilities of the Stable Diffusion (SD) model across different areas of its distribution using WordNet's taxonomy. To achieve this, we start with a general concept represented by a synset in WordNet, and recursively visit each node's hyponyms to create a hierarchy of related words. We limit our exploration to a maximum depth of 2 and consider only synsets with lemmas' frequency of occurrence above a chosen threshold for very dense hierarchies. Figure 7.2.1 presents an example of such a hierarchy.

To aid in the image generation process, we store the description for each synset-node. We use checkpoint 1.4 of Stable Diffusion as the basis for image generation and employ a pre-trained image feature extractor to generate image embeddings. We then conduct various experiments on these embeddings.

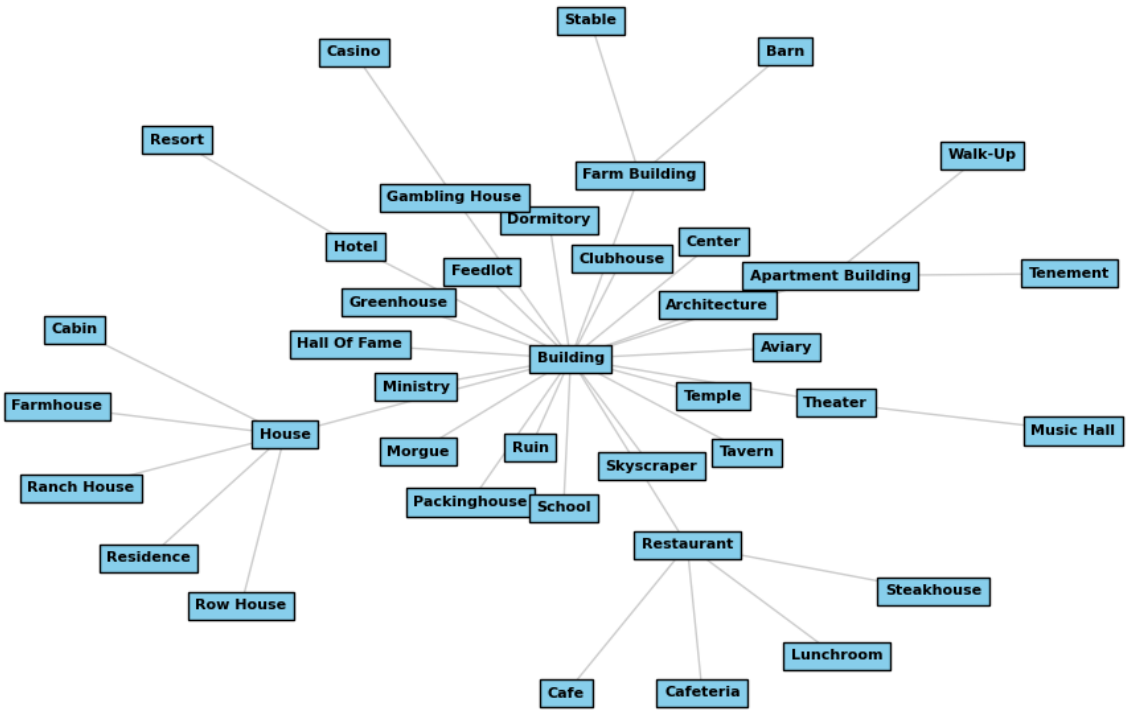


Figure 7.2.1: Example hyponym hierarchy generated from root synset "building"

7.2.1 Prompt and Image Generation with Stable Diffusion

This section of the thesis outlines the process of generating synthetic images using Stable Diffusion's text-conditional image generation. We used synsets retrieved from WordNet and concatenated each synset's name with its description to form a text prompt. We found that appending the word "photograph" at the end of the prompt improved the quality of the resulting images. However, we also observed that the synset descriptions could sometimes be vague or misleading, leading to poor quality images. To overcome this limitation, we employed Composable-Diffusion [30], a framework that allows for the combination of multiple prompts by assigning weights to each of them. In our case, we combined the original prompt with another version that only included the synset's name.

To determine the image generation parameters, we performed a grid-like search over different sampling methods, the number of sampling steps, and the classifier free guidance (CFG) scale. The CFG scale controls how strongly the image should conform to the prompt, affecting the variability of the resulting images. To ensure consistency and eliminate randomness, we set a constant seed during the search process. The final parameter values used for image generation were the following:

Parameter	**Value**
Sampler	LMS
Sampling Steps	35
CFG Scale	5.5

Once the synthetic images were generated, we utilized VGG16 [16], a convolutional neural network that was pre-trained on ImageNet [17], to extract image features in the form of low-dimensional vector representations. These vector representations were used for subsequent visualizations and experiments. It's worth noting that in this chapter, we will be using the terms "image" and "image feature vector" interchangeably since the feature vector captures the essential information in the image, and is a compact representation of the original image.

7.2.2 Classification to Hypernyms

In this experiment, we aimed to test whether the semantic content of the generated images adhered to the hierarchy dictated by their corresponding synsets. Specifically, we hypothesized that the distribution of images generated from a synset should be a superset of the distribution of images generated from its hyponyms.

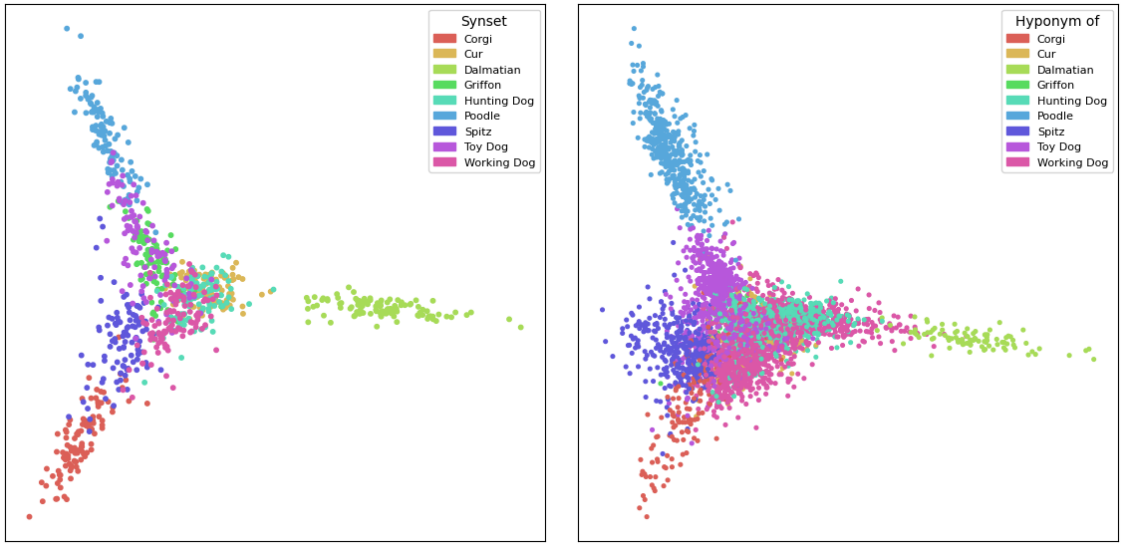
To quantify this property, we first trained a classifier on a dataset of images D_s generated from a set of synsets T_s that have at least one hyponym in our WordNet hierarchy. The classifier was trained to predict each image's corresponding synset. Next, we evaluated the same classifier on a dataset of images D_h generated from the set of hyponyms of T_s's synsets (T_h). A high classification accuracy would indicate that the distributions of each synset and its hyponyms were accurately represented in the image space of Stable Diffusion.

For visualizations, like the on in Figure 7.2.2, we utilized dimensionality reduction algorithms such as PCA and t-SNE to project the images in 2D space.

To efficiently test multiple model pipelines and select the optimal hyperparameters, we utilized auto-sklearn [31], an automated machine learning (AutoML) framework. Auto-sklearn uses Bayesian optimization methods to explore the hyperparameter space and various model architectures while taking into account past performance on similar datasets. The framework also constructs an ensemble of the highest-ranking models for improved performance.

7.2.3 Clustering Hyponyms

The objective of this experiment was to assess whether the images generated from semantically similar concepts remained close to each other in the image space. To achieve this, we performed unsupervised clustering on the dataset of images D_h generated from the hyponyms of T_s's synsets or T_h. We then compared the learned clusters with the true clusters, which were formed by the hypernyms of D_h's images.



(a) D_s (b) D_h

Figure 7.2.2: Distribution of images generated from synsets of T_s (left) and T_h (right). The WordNet graph used in this case was generated from root "dog".

To measure the similarity between the predicted and true clusterings, we utilized the adjusted rand index, which computes a similarity measure between two clusterings by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clusterings.

We performed a grid search to identify the optimal clustering algorithm and hyperparameter values. In addition, we used under sampling to address any class imbalances that may have existed in D_h, which was used for training.

In the case of hierarchical clustering, specifically agglomerative clustering, we obtained a hierarchy of image clusters. This hierarchy could be compared with the original hierarchy from WordNet, providing insights into the model's ability to generate semantically meaningful clusters of images.

Chapter 8

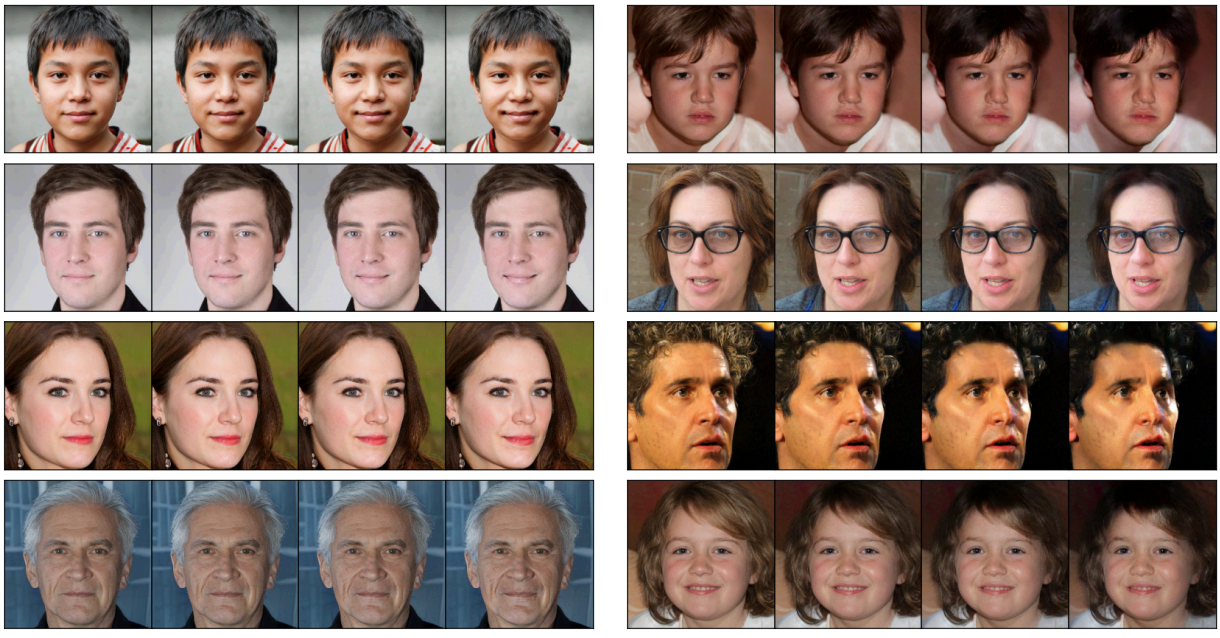
Results

8.1 Glow: Latent Space Exploration

This section presents a comparison between the proposed latent direction discovery method and established methods, based on both qualitative and quantitative results obtained through image manipulations using the discovered latent directions in Glow. The focus of the analysis is on the effectiveness and robustness of the latent directions, which are tested by gradually increasing the intensity of the manipulation and observing the corresponding changes in the images.

8.1.1 Qualitative Results

The manipulation results presented in Figures 8.1.1 and 8.1.3 results show that, while some of the images appear realistic, several become deformed at large manipulation intensities, revealing the limitations of the method. Additionally, the attribute entanglement becomes apparent at this scale, as illustrated by the correlation between the "Male" and "Beard" attributes in Figure 8.1.3.



(a) Smile (b) Dark Hair

Figure 8.1.1: Manipulations of different discovered facial attributes

To establish a reference point for comparison, this study implements the method presented in the Glow paper by Kingma et al. [8], which we refer to as GlowCeleb. In GlowCeleb, D_{source} and D_{target} are populated with images from CelebA-HQ, a dataset that provides facial attribute scores for its images. This approach ensures the quality of the two datasets but limits the possible manipulations to the attribute labels used in CelebA-HQ.

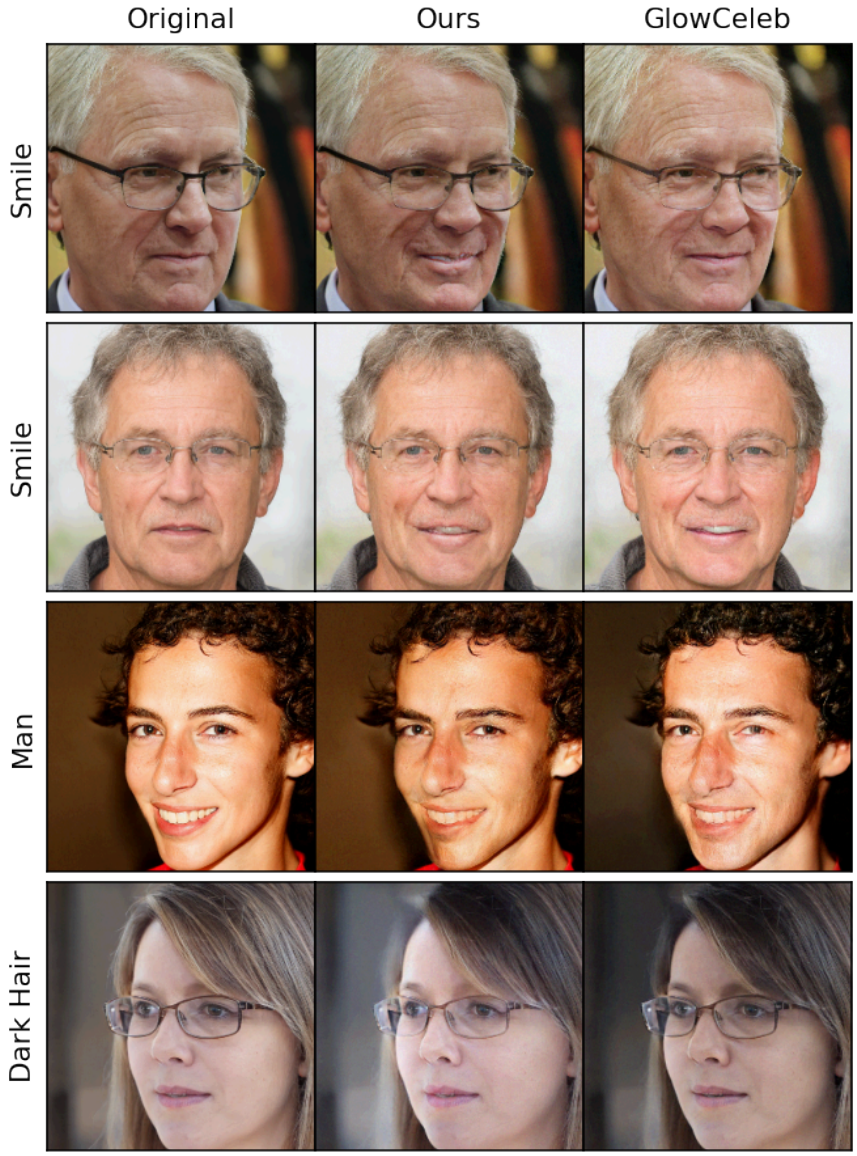


Figure 8.1.2: Comparison of our direction discovery method with the one used by Kingma in [8]

Despite the potential advantages of GlowCeleb, such as the use of labeled data to determine the corresponding latent direction, the proposed framework offers comparable manipulation quality while providing additional flexibility since the direction discovery is entirely dependent on the input text query. Its effectiveness is highly dependent on the quality and capacity of the text-conditional generative model employed and as these types of generative model continue to evolve and improve, the proposed framework's effectiveness is expected to increase correspondingly.

8.1.2 Quantitative Results

In this section, we present a quantitative evaluation of our model by computing and comparing the Manipulation Disentanglement Score (MDS) defined in Section 7.1.3 for different queries and across different latent direction discovery methods.

Table 8.1: MDS using CelebA-HQ facial attribute classifiers. The generative model whose latent space was used for the manipulations and the dataset it was trained on is also included. Higher MDS is better.

Method	Model	Dataset	Manipulation Disentanglement Score ↑				
			Smile	Young	Dark Hair	Gender	Overall
DisCo	StyleGAN2	FFHQ	0.68	0.51	-	-	0.60
GANSpace	StyleGAN2	FFHQ	0.24	-	0.54	0.84	0.54
InterfaceGAN	Progressive-GAN	CelebA-HQ	0.85	-	0.88	0.88	0.87
SGF	Progressive-GAN	CelebA-HQ	0.90	-	0.90	0.88	0.89
GlowCeleb	Glow	CelebA-HQ	0.66	0.88	0.75	0.88	0.79
Ours	Glow	CelebA-HQ	0.60	0.89	0.58	0.95	0.75

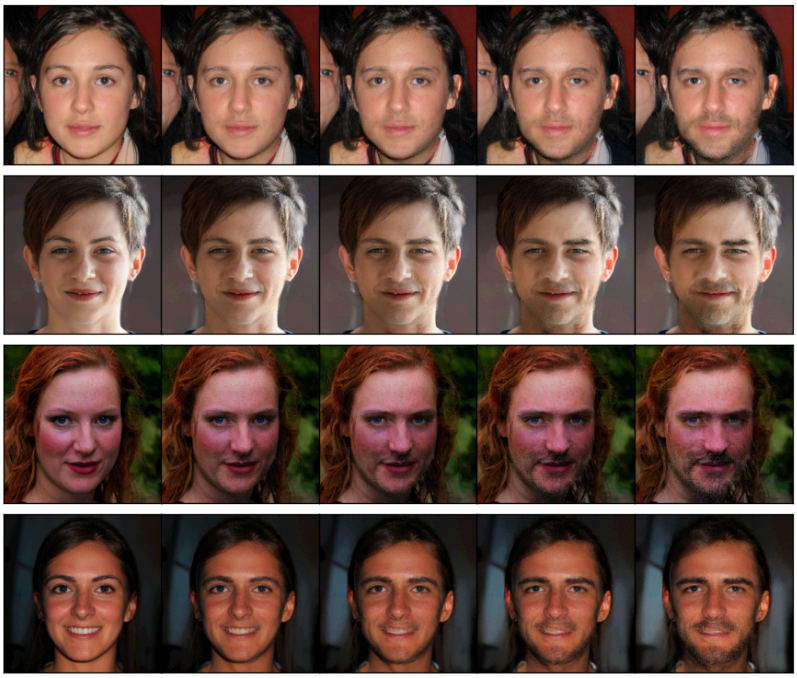
Table 8.1 presents the MDS for four facial attributes. These specific attributes were chosen because Ren [18] and Li [15] had reported their results on them. We observed that the MDS varied greatly depending on the examined attribute. Our method performed better on larger scale, structural facial manipulations but weaker on more detailed ones.

In addition to the MDS, we also computed the manipulation disentanglement curves (MDCs) for some of the discovered latent directions, as shown in Figure 8.1.4. We remind the reader that the MDS is defined as the area under the curve (AUC) of the MDC.

For GlowCeleb and our method, we also computed an alternate version of the MDS and MSC, which uses image similarity instead of a pre-trained attribute classifier to estimate the manipulation's disentanglement. To compute image similarity we extracted image features and calculated the cosine similarity between images as described in Section 7.1.3. Table 8.2 presents the MDS using image similarity. We observed that the MDSs using image similarity were lower than those computed using classifier-based manipulation disentanglement. However, relative to GlowCeleb, our method showed similar results.

Table 8.2: MDS using Image Similarity

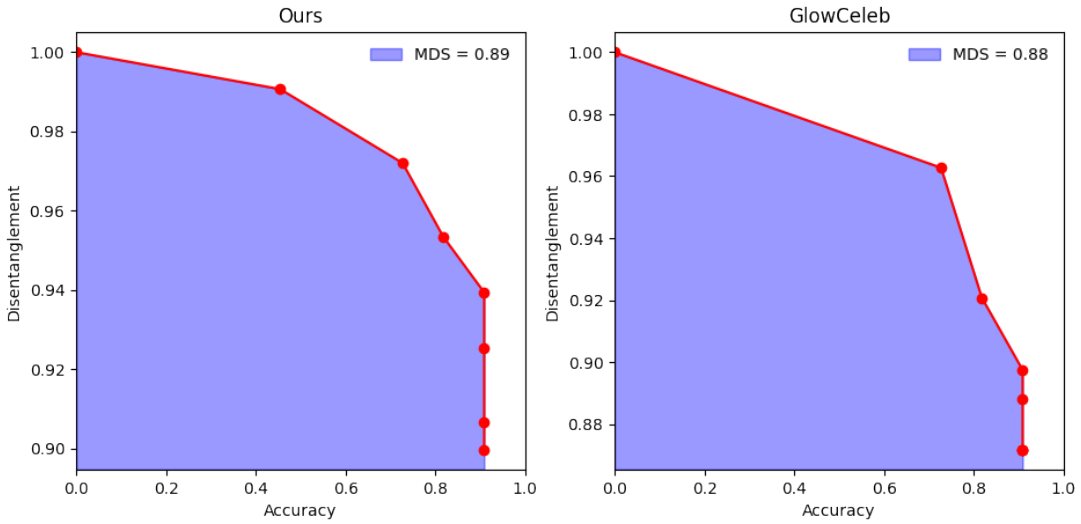
Method	Manipulation Disentanglement Score ↑				
	Smile	Young	Dark Hair	Gender	Overall
GlowCeleb	0.62	0.86	0.62	0.77	0.71
Ours	0.58	0.85	0.50	0.82	0.69



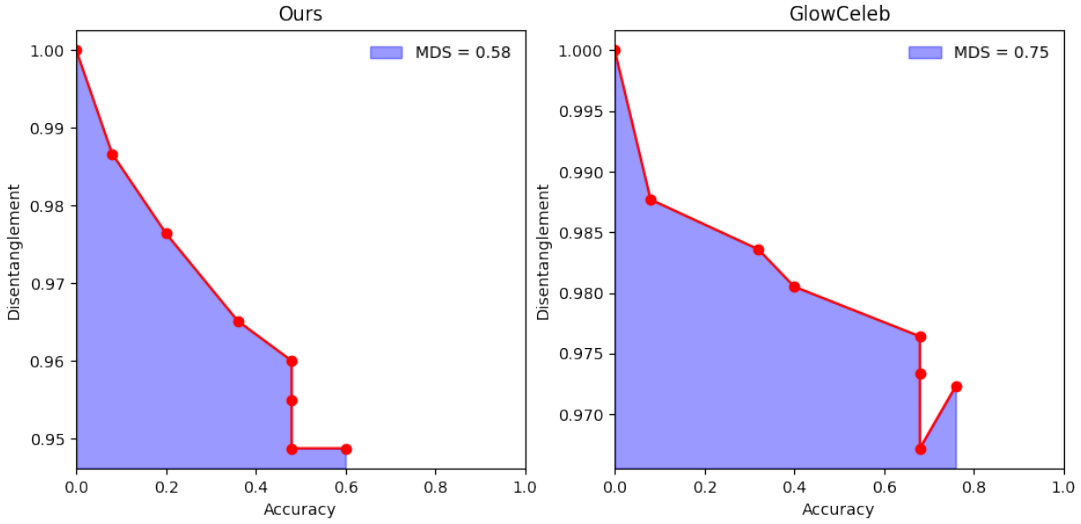
(a) Man

(b) Blonde Hair

Figure 8.1.3: Manipulations of different facial attributes at high intensity values



(a) Young



(b) Dark Hair

Figure 8.1.4: Manipulation Disentanglement Curves which visualize the robustness of the manipulation at high intensitities

8.2 Stable Diffusion Analysis

We mainly experimented on the graph that was generated from the "dog" synset depicted in Figure 8.2.1. Some sample images generated with SD for this hierarchy are also presented in Figure 8.2.2

We proceed to perform feature extraction using VGG16. We apply PCA on the extracted image features in order to reduce their dimensions and visualize them.

8.2.1 Classification Results

We utilized auto-sklearn for optimization. The highest performing model was an ensemble of classifiers consisting of Linear Discriminant Analysis, Extra Trees, Passive Aggressive, and Random Forest Classifiers. The classifier was trained on the image dataset D_s which contained all images generated from synsets that have at least one hyponym (T_s) in our chosen WordNet hierarchy.

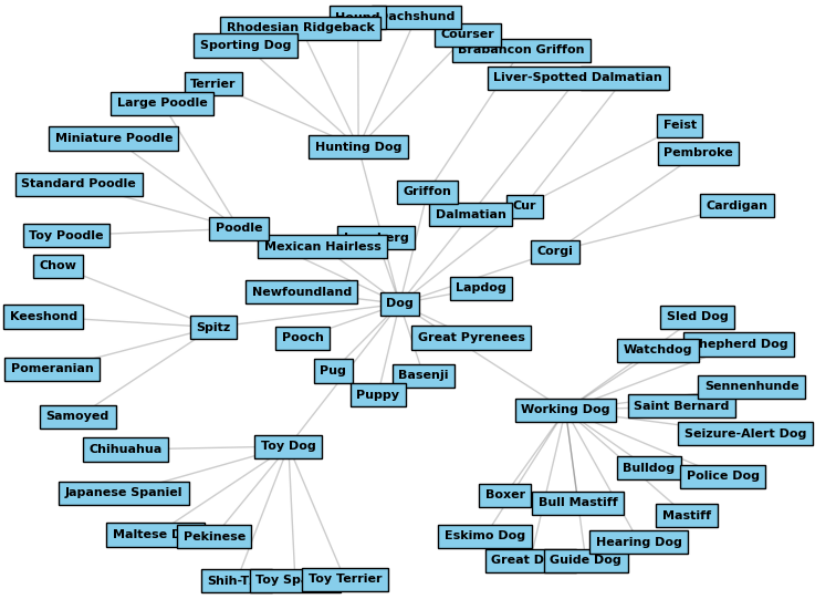


Figure 8.2.1: Hyponym hierarchy generated from root synset "dog"

Afterwards, it was evaluated on the image dataset D_h which contains all images that were derived from all the hyponyms of T_s's synsets. The resulting classification metrics are presented in Table 8.3 and are interpreted below. Additionally, the classifier's predictions are visualized in Figure 8.2.3.

Class	**Dataset D_s**			**Dataset D_h**		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Corgi	1.00	0.92	0.96	0.97	0.62	0.75
Cur	0.90	1.00	0.95	0.04	0.14	0.06
Dalmatian	1.00	1.00	1.00	0.95	1.00	0.97
Griffon	0.95	0.88	0.91	0.05	0.18	0.08
Hunting Dog	0.72	0.78	0.75	0.53	0.33	0.41
Poodle	0.97	0.91	0.94	0.99	0.98	0.99
Spitz	0.94	0.94	0.94	0.59	0.99	0.74
Toy Dog	0.85	0.89	0.87	0.75	0.66	0.70
Working Dog	0.71	0.74	0.73	0.80	0.44	0.57
Accuracy	-	-	0.89	-	-	0.57
Macro Avg.	0.89	0.90	0.89	0.63	0.59	0.59
Weighted Avg.	0.90	0.89	0.89	0.71	0.57	0.61

Table 8.3: Classification metrics for datasets D_s and D_h. The classifier was trained on a subset of D_s.

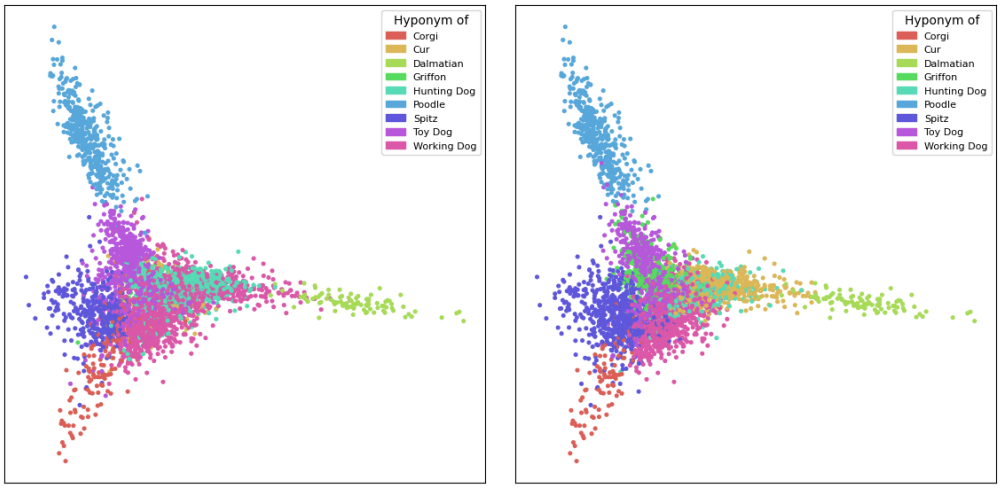
We based our conclusions primarily on the weighted averages of the metrics, given the class imbalance present in the test set D_h. Our analysis indicates that the classifier performs exceptionally well in modeling the distribution of D_s. However, the classifier's scores on D_h, which represents hyponym generated images, are not as high. This implies that for certain synset and hyponym pairs, the synset images distribution is not a superset of the hyponym images distribution.

To approximate the degree to which the image distribution of a synset contains the image distribution of its hyponyms, we can use the recall score achieved by the classifier on the test set D_h for each synset. This is because the recall score on D_h quantifies the percentage of hyponym images that were correctly classified

Figure 8.2.2: Images generated from synsets derived from root synset "dog"

to their hypernym synset. If the image distribution of a synset includes the image distribution of all of its hyponyms, then we would expect a recall score of 1 on this particular synset class on the test set D_h.

The recall metric's weighted average on the test set D_h is 0.57, indicating that the image distribution derived from a synset, on average, encompasses 57% of the image distribution derived from the synset's hyponyms. This metric can be viewed as an estimation of how effectively Stable Diffusion encodes WordNet's hierarchy for this specific subgraph derived from the root "dog."



(a) True Synset

(b) Predicted Synset

Figure 8.2.3: Classification of images of D_h into their hypernyms in our WordNet hierarchy

8.2.2 Clustering Results

For the clustering task, we conducted a grid search to determine the optimal feature preprocessing pipeline and hyperparameters for the agglomerative clustering algorithm. Through this process, we were able to develop a model that achieved an adjusted rand index score of 0.45, indicating a moderately strong correlation between the clusters and the synsets they contain.

To facilitate interpretation and analysis of the resulting clusters, we named each cluster after the most frequent synset it contained. This approach allows us to quickly identify the main themes or categories represented in each cluster, providing insights into the underlying patterns and relationships between the images.

We visualize the dendrogram that corresponds to the hierarchical clustering process in Figure 8.2.4. The names of the leaves are of the form Synset/HyponymSynset. However, the clustering was derived by training only on hyponym images (D_h). We include each hyponym's hypernym to visually demonstrate that the algorithm can successfully group the hyponym clusters into hypernym clusters.

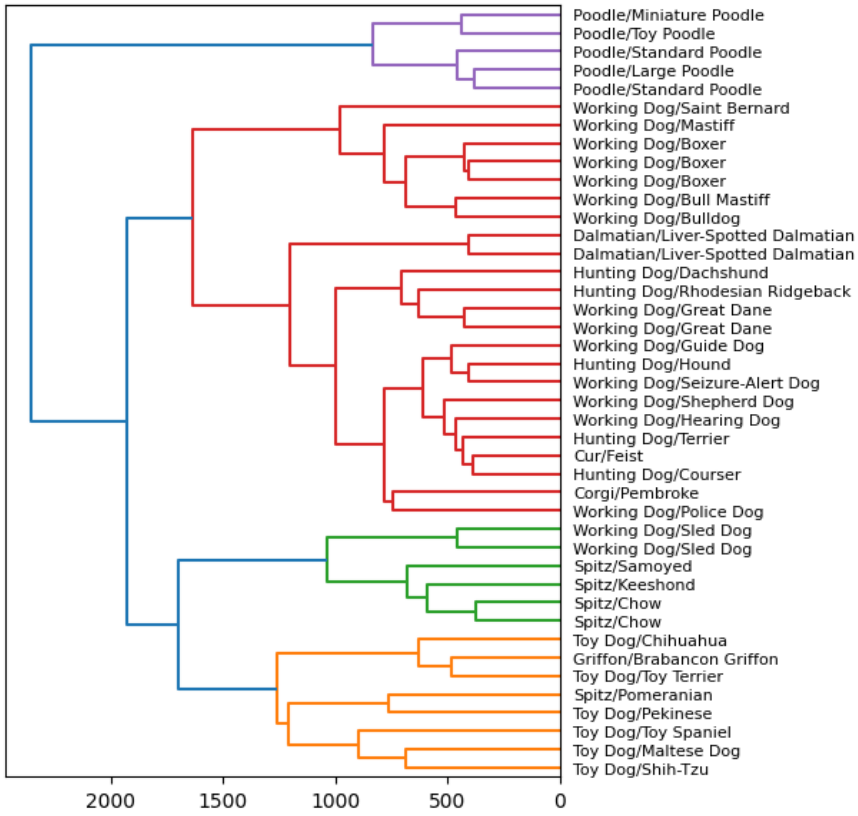


Figure 8.2.4: Hierarchical Clustering Dendrogram

Chapter 9

Conclusion

9.0.1 Discussion

In this work, we proposed a novel unsupervised method for latent direction discovery in the latent space of Glow, which is entirely text-conditional. Our method leverages a textual description to discover a latent direction that controls the described semantic attribute of the image. To our knowledge, this is the first work that follows a text-guided approach for the Glow model. Prior works utilized other methods to find latent directions. Unsupervised methods typically lack the ability to automatically name directions and require user annotations on the discovered directions. Moreover, they do not provide insights regarding the properties of the latent space. Our text-guided method bypasses these issues and achieves results comparable to other state-of-the-art direction discovery methods. These findings are supported by a quantitative comparison using a robust, model agnostic metric called manipulation disentanglement score.

Additionally, we investigated the variability of Stable Diffusion's text-conditional image generation capabilities across different areas of its sample distribution. We utilized WordNet's hierarchical knowledge to explore how closely related textual concepts and their differences are represented in the image space, and quantified these qualities. Our experiments showed that for a subset of WordNet, Stable Diffusion managed to produce diverse results when fed closely related concepts, as evidenced by the achieved classification scores on the generated images. However, we found that WordNet's hierarchy is not always modeled correctly by Stable Diffusion. For a number of synset and hyponym pairs, the image generations derived from the synset were not necessarily more general than the image generations of its hyponym, indicating the existence of a bias towards certain hyponyms and away from others. We measured the level of bias for a subgraph of WordNet derived from the root "dog," but the execution of the pipeline for different concepts is straightforward.

9.0.2 Future Directions

In this section, we suggest a few possible directions to further improve on this work. Regarding our latent direction discovery method, we could explore the use of text-conditional image generation models other than StyleCLIP. For example, incorporating diffusion models for image manipulation as part of our proposed pipeline could yield interesting results. Additionally, while the textual descriptions used in this work were relatively simple, our method's text-conditional nature allows for the use of more detailed inputs that describe more complex attributes. Therefore, evaluating the performance of our proposed method on such complex inputs could provide insight into its limitations. Furthermore, we can improve the proposed method by automatically choosing the parameters of StyleCLIP's Global Direction algorithm. Specifically, we can select the optimal set of parameters that maximizes the manipulation disentanglement score of the resulting latent directions.

Expanding on the future directions for the evaluation of Stable Diffusion, we can investigate the effectiveness of the model on larger subsets of WordNet to gain a more comprehensive understanding of its performance. This could involve analyzing the model's ability to generate diverse and high-quality images for a variety of synsets and hyponyms belonging to subsets of WordNet, and comparing its results with those of other

text-guided diffusion models. Furthermore, to address the observed bias towards certain concepts detected in this work, we can investigate different strategies for improving the training process of Stable Diffusion. For instance, we can balance the distribution of concepts in the training data or use more sophisticated loss functions that encourage diversity and fairness.

Bibliography

[1] Goodfellow, I. et al. "Generative adversarial networks". In: *Communications of the ACM* 63.11 (2020), pp. 139–144.

[2] Kingma, D. P. and Welling, M. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

[3] Croitoru, F.-A. et al. "Diffusion models in vision: A survey". In: *arXiv preprint arXiv:2209.04747* (2022).

[4] Rombach, R. et al. "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10684–10695.

[5] Saharia, C. et al. "Photorealistic text-to-image diffusion models with deep language understanding". In: *arXiv preprint arXiv:2205.11487* (2022).

[6] Ruiz, N. et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation". In: *arXiv preprint arXiv:2208.12242* (2022).

[7] Ramesh, A. et al. "Zero-shot text-to-image generation". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 8821–8831.

[8] Kingma, D. P. and Dhariwal, P. "Glow: Generative flow with invertible 1x1 convolutions". In: *Advances in neural information processing systems* 31 (2018).

[9] Rezende, D. and Mohamed, S. "Variational inference with normalizing flows". In: *International conference on machine learning*. PMLR. 2015, pp. 1530–1538.

[10] Dinh, L., Krueger, D., and Bengio, Y. "Nice: Non-linear independent components estimation". In: *arXiv preprint arXiv:1410.8516* (2014).

[11] Dinh, L., Sohl-Dickstein, J., and Bengio, S. "Density estimation using real nvp". In: *arXiv preprint arXiv:1605.08803* (2016).

[12] Patashnik, O. et al. "Styleclip: Text-driven manipulation of stylegan imagery". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 2085–2094.

[13] Radford, A. et al. "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

[14] Karras, T. et al. "Analyzing and improving the image quality of stylegan". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 8110–8119.

[15] Li, M., Jin, Y., and Zhu, H. "Surrogate gradient field for latent space manipulation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 6529–6538.

[16] Simonyan, K. and Zisserman, A. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[17] Deng, J. et al. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[18] Ren, X. et al. "Learning disentangled representation by exploiting pretrained generative models: A contrastive learning view". In: *International Conference on Learning Representations*. 2021.

[19] Miller, G. A. "WordNet: a lexical database for English". In: *Communications of the ACM* 38.11 (1995), pp. 39–41.

[20] Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning*. MIT press, 2016.

[21] Reynolds, D. A. et al. "Gaussian mixture models." In: *Encyclopedia of biometrics* 741.659-663 (2009).

[22] Izenman, A. J. "Review papers: Recent developments in nonparametric density estimation". In: *Journal of the american statistical association* 86.413 (1991), pp. 205–224.

[23] Härkönen, E. et al. "Ganspace: Discovering interpretable gan controls". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 9841–9850.

[24] Voynov, A. and Babenko, A. “Unsupervised discovery of interpretable directions in the gan latent space”. In: *International conference on machine learning*. PMLR. 2020, pp. 9786–9796.

[25] Shen, Y. et al. “Interfacegan: Interpreting the disentangled face representation learned by gans”. In: *IEEE transactions on pattern analysis and machine intelligence* 44.4 (2020), pp. 2004–2018.

[26] Yang, H. et al. “Discovering interpretable latent space directions of gans beyond binary attributes”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 12177–12185.

[27] Wu, Z., Lischinski, D., and Shechtman, E. “Stylespace analysis: Disentangled controls for stylegan image generation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 12863–12872.

[28] Szegedy, C. et al. “Inception-v4, inception-resnet and the impact of residual connections on learning”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31. 1. 2017.

[29] Cao, Q. et al. “Vggface2: A dataset for recognising faces across pose and age”. In: *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE. 2018, pp. 67–74.

[30] Liu, N. et al. “Compositional visual generation with composable diffusion models”. In: *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*. Springer. 2022, pp. 423–439.

[31] Feurer, M. et al. “Efficient and robust automated machine learning”. In: *Advances in neural information processing systems* 28 (2015).

やった!