



## Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών  
Εργαστήριο Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης

Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών  
Επιστήμη Δεδομένων και Μηχανική Μάθηση

---

**Καθοδηγούμενη από Σκίτσα και Κειμενικές Περιγραφές  
Σύνθεση Εικόνων με Χρήση Μοντέλων Διάχυσης**

---

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

**Ηλία Μήτσουρα**

**Επιβλέπων:** Αθανάσιος Βουλόδημος  
Επίκουρος Καθηγητής Ε.Μ.Π.

**Συνεπιβλέπουσα:** Παρασκευή Τζούβελη  
Μέλος Ε.ΔΙ.Π. Ε.Μ.Π.

Αθήνα, Μάρτιος 2024





## Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών  
Εργαστήριο Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης

Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών  
Επιστήμη Δεδομένων και Μηχανική Μάθηση

---

**Καθοδηγούμενη από Σκίτσα και Κειμενικές Περιγραφές  
Σύνθεση Εικόνων με Χρήση Μοντέλων Διάχυσης**

---

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

**Ηλία Μήτσουρα**

**Επιβλέπων:** Αθανάσιος Βουλόδημος  
Επίκουρος Καθηγητής Ε.Μ.Π.

**Συνεπιβλέπουσα:** Παρασκευή Τζούβελη  
Μέλος Ε.ΔΙ.Π. Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 29<sup>η</sup> Μαρτίου 2024.

.....  
Αθανάσιος Βουλόδημος  
Επίκουρος Καθηγητής  
Ε.Μ.Π.

.....  
Ανδρέας-Γεώργιος Σταφυλοπάτης  
Καθηγητής  
Ε.Μ.Π.

.....  
Γεώργιος Στάμου  
Καθηγητής  
Ε.Μ.Π.

Αθήνα, Μάρτιος 2024

.....  
**Ηλίας Μήτσουρας**

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

© Ηλίας Μήτσουρας, 2024. Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

## Περίληψη

---

Η καθοδηγούμενη από σκίτσα και κειμενικές περιγραφές σύνθεση εικόνων αποσκοπεί στην παραγωγή ρεαλιστικών και υψηλής πιστότητας εικόνων, οι οποίες αφενός συμμορφώνονται με το νοηματικό περιεχόμενο των κειμενικών περιγραφών και αφετέρου ακολουθούν πιστά τα σκίτσα αναφοράς ως προς τα χωρικά περιγράμματα. Τα τελευταία χρόνια, τα μοντέλα διάχυσης έχουν επιδείξει θεαματικά αποτελέσματα σε προβλήματα σύνθεσης, παράγοντας ρεαλιστικές και υψηλής ανάλυσης εικόνες. Παρά τη μεγάλη τους αυτή επιτυχία, υστερούν ακόμη σημαντικά στο προαναφερθέν πρόβλημα της καθοδηγούμενης από σκίτσα σύνθεσης εικόνων, κατά το οποίο καλούνται να γεφυρώσουν το κενό μεταξύ της αφαιρετικής και αδόμητης φύσης των ελεύθερων σκίτσων και του υψηλού επιπέδου των λεπτομερειών των πραγματικών εικόνων. Μία πρόσφατη προσέγγιση βασίζεται στην υλοποίηση ενός MLP latent edge predictor, ο οποίος εκπαιδεύεται ώστε να προβλέπει τον χάρτη ακμών της παραγόμενης εικόνας σε κάθε βήμα της διαδικασίας αποθορυβοποίησης. Ο χάρτης αυτός αξιοποιείται εν συνεχεία για την καθοδήγηση των χωρικών περιγραμμάτων της εικόνας. Παρά τα σχετικά ικανοποιητικά αποτελέσματα που παρουσιάζει η εν λόγω μέθοδος, δε λαμβάνει υπόψιν τις χωρικές συσχετίσεις των pixels, ενώ συγχρόνως, απαιτεί ένα σημαντικό αριθμό βημάτων κατά τη διαδικασία της αποθορυβοποίησης, με αποτέλεσμα να καθίσταται ιδιαίτερος χρονοβόρα. Για την αντιμετώπιση των ανωτέρω περιορισμών, στα πλαίσια της παρούσας εργασίας προτείνεται ένα πλαίσιο καθοδήγησης, το οποίο βασίζεται στη χρήση ενός U-Net latent edge predictor, ο οποίος, λόγω της συνελκτικής του φύσης, είναι ικανός να αποτυπώνει αποτελεσματικά, τόσο τοπικά όσο και καθολικά χαρακτηριστικά, αντιμετωπίζοντας τις εισόδους ενιαία και όχι pixel-wise. Επιπρόσθετα, για την ενίσχυση της σθεναρότητας της όλης διαδικασίας, στο προτεινόμενο πλαίσιο προστίθεται και ένα δίκτυο απλοποίησης σκίτσων, το οποίο προσφέρει στο χρήστη τη δυνατότητα απλοποίησης των σκίτσων εισόδου. Τα πειραματικά αποτελέσματα, σε συνδυασμό με τη γνώμη των χρηστών, αποδεικνύουν ότι η χρήση του προτεινόμενου U-Net latent edge predictor οδηγεί σε πιο ρεαλιστικές εικόνες, οι οποίες είναι καλύτερα ευθυγραμμισμένες με τα χωρικά περιγράμματα των σκίτσων αναφοράς, ενώ συγχρόνως, μειώνει δραστικά τα απαιτούμενα βήματα αποθορυβοποίησης και κατά συνέπεια το συνολικό χρόνο εκτέλεσης.

**Λέξεις-Κλειδιά:** Μοντέλα Διάχυσης · Καθοδηγούμενη από Σκίτσα Σύνθεση Εικόνων · U-Net latent edge predictor · Όραση Υπολογιστών



# Abstract

---

Sketch-guided text-to-image synthesis aims to obtain realistic and high fidelity images that adhere to the semantic content of the textual descriptions, while faithfully following the spatial outlines of the corresponding sketches. In recent years, diffusion models have demonstrated remarkable results, producing realistic and high-resolution images and thus exhibiting a clear superiority over GANs in text-to-image synthesis tasks. Despite their significant success, they still fall behind in the aforementioned task of sketch-guided image synthesis, as they try to bridge the gap between the abstract and schematic nature of freehand sketches and the rich details of real-world images. A recent approach tries to address this task by employing a per-pixel MLP latent edge predictor to predict the edge map of the generated image at each step of the inverse diffusion process. This edge map is then used to guide the image's spatial outlines towards the reference sketch. Despite yielding relatively satisfactory results, this method does not take into account spatial correlations between pixels and demands numerous denoising iterations to produce satisfying images, leading to time inefficiency. To overcome these limitations, we propose a framework that utilizes a U-Net latent edge predictor, which due to its convolutional nature is capable of effectively capturing both local and global features, treating inputs as whole rather than in a pixel-wise manner. Moreover, the proposed guidance framework is enhanced with the addition of a sketch simplification network, which offers the user the ability to preprocess and simplify input sketches. Experimental results in conjunction with user feedback show that the use of the proposed U-Net latent edge predictor leads to more realistic results, that are better aligned with the spatial outlines of the reference sketches, while significantly reducing the number of required denoising steps and consequently, the overall execution time.

**Keywords:** Diffusion Models · Sketch-guided Text-to-Image Synthesis · U-Net latent edge predictor · Computer Vision





## Ευχαριστίες

---

Η παρούσα διπλωματική εργασία σηματοδοτεί το τέλος των μεταπτυχιακών σπουδών μου στο διατμηματικό πρόγραμμα της Επιστήμης Δεδομένων και Μηχανικής Μάθησης, της σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου. Οι γνώσεις και η εμπειρία που απέκτησα κατά τη διάρκεια των σπουδών μου στα πεδία της Μηχανικής Μάθησης και της Τεχνητής Νοημοσύνης, συνέβαλαν καθοριστικά στην ενίσχυση του επιστημονικού μου υποβάθρου και στη διεύρυνση των ερευνητικών μου οριζώντων. Στο πλαίσιο αυτό, θα ήθελα να ευχαριστήσω όλους όσους με στήριξαν και συνέβαλαν στην επιτυχή περάτωση της πορείας αυτής.

Πρωτίστως, θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου κύριο Αθανάσιο Βουλόδημο, για την ευκαιρία που μου προσέφερε να εργαστώ πάνω στο παρόν θέμα, στο εργαστήριο Τεχνητής Νοημοσύνης και Συστημάτων Μάθησης. Έπειτα, θα ήθελα να ευχαριστήσω ιδιαίτερα την κυρία Παρασκευή Τζούβελη, για την άριστη συνεργασία που είχαμε καθόλη τη διάρκεια εκπόνησης της διπλωματικής και την υποστήριξή της. Ακολούθως, θερμές ευχαριστίες θα ήθελα να απευθύνω στον Ελευθέριο Τσώνη, για την εξαιρετική αμεσότητα και προθυμία του στη συζήτηση των όποιων σκέψεων και ερωτημάτων είχα καθόλη τη διάρκεια της εκπόνησης της διπλωματικής.

Τέλος, ιδιαίτερες ευχαριστίες εκφράζω στους φίλους μου και κυρίως στα αδέρφια και τους γονείς μου, Αθανάσιο και Βασιλική, χάρη στους οποίους έγινα ο άνθρωπος που είμαι σήμερα και χωρίς τους οποίους, όλη αυτή η πορεία δεν θα ήταν εφικτή.



# Περιεχόμενα

---

|   |             |
|---|-------------|
| <b>Κατάλογος Σχημάτων</b>   | <b>xi</b>   |
| <b>Κατάλογος Πινάκων</b>  | <b>xiii</b> |
| <b>I Εισαγωγή</b>   | <b>1</b>    |
| <b>1 Περιγραφή Προβλήματος</b>                                      | <b>3</b>    |
| 1.1 Δομή Εργασίας . . . . .   | 4           |
| <b>II Θεωρητικό Υπόβαθρο</b>  | <b>7</b>    |
| <b>2 Diffusion Models</b>   | <b>9</b>    |
| 2.1 Denoising Diffusion Probabilistic Models . . . . .              | 9           |
| 2.2 Score-based Generative Models . . . . .                         | 16          |
| 2.2.1 Score Matching . . . . .                                      | 16          |
| 2.2.2 Sliced Score Matching . . . . .                               | 17          |
| 2.2.3 Denoising Score Matching . . . . .                            | 19          |
| 2.2.4 Langevin Dynamics . . . . .                                   | 19          |
| 2.2.5 Noise Conditional Score Networks (NCSN) . . . . .             | 19          |
| 2.3 Score-based Generative Modeling μέσω SDEs . . . . .             | 21          |
| 2.3.1 SMLD ως Διακριτοποίηση Variance Exploding (VE) SDE . . . . .  | 22          |
| 2.3.2 DDPM ως Διακριτοποίηση Variance Preserving (VP) SDE . . . . . | 23          |
| 2.3.3 Επίλυση της Αντίστροφης SDE . . . . .                         | 24          |
| 2.4 Denoising Diffusion Implicit Models . . . . .                   | 26          |
| 2.4.1 Επιτάχυνση Διαδικασίας Παραγωγής Δειγμάτων . . . . .          | 28          |
| <b>3 Text-to-Image Synthesis</b>                                    | <b>30</b>   |
| 3.1 Καθοδήγηση Μοντέλων Διάχυσης . . . . .                          | 31          |
| 3.1.1 Classifier Guidance . . . . .                                 | 31          |

|            |  |           |
|------------|--|-----------|
| 3.1.2      | Classifier-Free Guidance . . . . .                         | 36        |
| 3.2        | Stable Diffusion . . . . .                                 | 38        |
| 3.2.1      | Autoencoder (VAE) . . . . .                                | 38        |
| 3.2.2      | Denoising U-Net . . . . .                                  | 40        |
| <b>4</b>   | <b>Sketch-Guided Image Synthesis</b>                       | <b>45</b> |
| 4.1        | Latent Edge Predictor . . . . .                            | 47        |
| 4.2        | Παραγωγή νέων εικόνων . . . . .                            | 51        |
| 4.3        | Τροποποίηση Αρχιτεκτονικής Latent Edge Predictor . . . . . | 53        |
| 4.4        | Προσθήκη Δικτύου Απλοποίησης Σκίτσων . . . . .             | 56        |
| <b>III</b> | <b>Πειραματικό Μέρος</b>                                   | <b>59</b> |
|            | <b>Εισαγωγή</b>  | <b>60</b> |
| <b>5</b>   | <b>Κατασκευή Συνόλου Δεδομένων Εκπαίδευσης</b>             | <b>61</b> |
| <b>6</b>   | <b>Αποτελέσματα - per-pixel MLP</b>                        | <b>63</b> |
| 6.1        | Παράμετροι Εκπαίδευσης . . . . .                           | 63        |
| 6.2        | Παραδείγματα Σύνθεσης . . . . .                            | 64        |
| 6.2.1      | Επίδραση Αρχικού Θορύβου . . . . .                         | 69        |
| 6.2.2      | Επίδραση Sketch-guidance Strength ( $\beta$ ) . . . . .    | 70        |
| 6.2.3      | Επίδραση Πλήθους Βημάτων Αντίστροφης Διαδικασίας Διάχυσης  | 71        |
| 6.3        | Γενικά συμπεράσματα . . . . .                              | 73        |
| <b>7</b>   | <b>Αποτελέσματα - U-Net</b>                                | <b>74</b> |
| 7.1        | Παραδείγματα Σύνθεσης . . . . .                            | 75        |
| 7.1.1      | Επίδραση Αρχικού Θορύβου . . . . .                         | 78        |
| 7.1.2      | Επίδραση Sketch-guidance Strength ( $\beta$ ) . . . . .    | 79        |
| 7.1.3      | Επίδραση Πλήθους Βημάτων Αντίστροφης Διαδικασίας Διάχυσης  | 80        |
| 7.2        | Χρήση Δικτύου Απλοποίησης Σκίτσων . . . . .                | 82        |
| 7.3        | Γενικά συμπεράσματα . . . . .                              | 83        |
| <b>8</b>   | <b>Συγκριτική Αξιολόγηση</b>                               | <b>84</b> |
| 8.1        | Ποιοτική Αξιολόγηση . . . . .                              | 84        |

---

|           |  |            |
|-----------|--|------------|
| 8.2       | Ποσοτική Αξιολόγηση . . . . .                      | 88         |
| 8.2.1     | Recall . . . . .                                   | 88         |
| 8.2.2     | User Studies . . . . .                             | 90         |
| <b>9</b>  | <b>Συμπεράσματα και Μελλοντική Έρευνα</b>          | <b>92</b>  |
| <b>IV</b> | <b>Παραρτήματα</b>                                 | <b>94</b>  |
|           | <b>Παράρτημα Α. Επιπλέον Παραδείγματα Σύνθεσης</b> | <b>95</b>  |
|           | <b>Βιβλιογραφία</b>                                | <b>100</b> |



## Κατάλογος Σχημάτων

---

|            |  |    |
|------------|--|----|
| Σχήμα 2.1: | Διαδικασία διάχυσης σε DDPM . . . . .  | 10 |
| Σχήμα 2.2: | Λειτουργία Score-based Generative μοντέλων μέσω στοχαστικών διαφορικών εξισώσεων. . . . .  | 21 |
| Σχήμα 2.3: | Μη Μαρκοβιανές διαδικασίες . . . . .   | 27 |
| Σχήμα 3.1: | Αρχιτεκτονική Stable Diffusion [22]. . . . .   | 38 |
| Σχήμα 3.2: | Αρχιτεκτονική δικτύου attention U-Net. . . . .   | 42 |
| Σχήμα 3.3: | Δομή ResNet block της αρχιτεκτονικής του δικτύου U-Net του Σχ.3.2. . . . .   | 43 |
| Σχήμα 3.4: | Διαδικασία εκπαίδευσης conditional U-Net. . . . .  | 44 |
| Σχήμα 4.1: | Σχήμα εκπαίδευσης per-pixel MLP latent edge predictor [4]. . . . .   | 48 |
| Σχήμα 4.2: | Sketch-Guided Text-to-Image Synthesis [4]. . . . .   | 51 |
| Σχήμα 4.3: | Αρχιτεκτονική U-Net latent edge predictor. . . . .   | 54 |
| Σχήμα 4.4: | Αρχικά και απλοποιημένα σκίτσα με χρήση του sketch simplification network . . . . .  | 56 |
| Σχήμα 4.5: | Σχηματική αναπαράσταση συνολικής διαδικασίας καθοδηγούμενης από σκίτσα text-to-image σύνθεσης. . . . .   | 58 |
| Σχήμα 5.1: | Δείγματα τριπλετών $(x, e, y)$ του συνόλου των δεδομένων εκπαίδευσης. . . . .  | 62 |
| Σχήμα 6.1: | Παραδείγματα καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων, με χρήση του δικτύου per-pixel MLP ως latent edge predictor. Για όλα τα παραδείγματα χρησιμοποιούνται οι παράμετροι του Πίνακα 6.2 κατά τη διαδικασία της αντίστροφης διάχυσης. . . . . | 66 |
| Σχήμα 6.2: | Παραδείγματα καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων, με χρήση του δικτύου per-pixel MLP ως latent edge predictor. (συνέχεια) . . . . .   | 67 |
| Σχήμα 6.3: | Παραδείγματα καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων, με χρήση του δικτύου per-pixel MLP ως latent edge predictor. (συνέχεια) . . . . .   | 68 |
| Σχήμα 6.4: | Παραδείγματα καθοδηγούμενης από σκίτσα σύνθεσης εικόνων, με χρήση του per-pixel MLP latent edge predictor και διαφορετικές αρχικοποιήσεις θορύβου. . . . .   | 69 |

|            |   |    |
|------------|---|----|
| Σχήμα 6.5: | Παραδείγματα καθοδηγούμενης από σκίτσα σύνθεσης εικόνων, με χρήση του per-pixel MLP latent edge predictor και διαφορετικές τιμές της παραμέτρου $\beta$ . . . . .   | 71 |
| Σχήμα 6.6: | Παραδείγματα καθοδηγούμενης από σκίτσα σύνθεσης εικόνων, με χρήση του per-pixel MLP latent edge predictor και διαφορετικές τιμές αριθμού βημάτων αποθορυβοποίησης. . . . .  | 72 |
| Σχήμα 7.1: | Παραδείγματα καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων, με χρήση του δικτύου U-Net ως latent edge predictor. Για όλα τα παραδείγματα χρησιμοποιούνται οι παράμετροι του Πίνακα 7.2 κατά τη διαδικασία της αντίστροφης διάχυσης. . . . .  | 76 |
| Σχήμα 7.2: | Παραδείγματα καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων, με χρήση του δικτύου U-Net ως latent edge predictor. (συνέχεια) . . . . .  | 77 |
| Σχήμα 7.3: | Παραδείγματα καθοδηγούμενης από σκίτσα text-to-image σύνθεσης, με χρήση του U-Net latent edge predictor και διαφορετικές αρχικοποιήσεις θορύβου. . . . .  | 79 |
| Σχήμα 7.4: | Παραδείγματα καθοδηγούμενης από σκίτσα σύνθεσης εικόνων, με χρήση του U-Net latent edge predictor και διαφορετικές τιμές της παραμέτρου $\beta$ . . . . .   | 80 |
| Σχήμα 7.5: | Παραδείγματα καθοδηγούμενης από σκίτσα σύνθεσης εικόνων, με χρήση του U-Net latent edge predictor και διαφορετικές τιμές αριθμού βημάτων αποθορυβοποίησης. . . . .  | 81 |
| Σχήμα 7.6: | Παραδείγματα καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων. Για κάθε τριάδα παρουσιάζονται από αριστερά προς τα δεξιά: το σκίτσο αναφοράς και η αντίστοιχη κειμενική περιγραφή, η εικόνα που παράγεται χωρίς και με χρήση του δικτύου απλοποίησης σκίτσων. . . . .   | 82 |
| Σχήμα 8.1: | Παραδείγματα καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων με χρήση αμφότερων των U-Net και per-pixel MLP latent edge predictors. Σε κάθε τριάδα παραδειγμάτων παρατίθενται από αριστερά προς τα δεξιά: το σκίτσο αναφοράς και η κειμενική περιγραφή, η παραγόμενη εικόνα με χρήση του U-Net και η παραγόμενη εικόνα με χρήση του MLP latent edge predictor. . . . . | 86 |
| Σχήμα 8.2: | Παραδείγματα καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων με χρήση αμφότερων των U-Net και per-pixel MLP latent edge predictors. (συνέχεια) . . . . .   | 87 |
| Σχήμα 8.3: | Διαδικασία υπολογισμού μετρικής Recall μεταξύ σκίτσου αναφοράς και εξαγόμενου χάρτη ακμών της παραγόμενης εικόνας. . . . .  | 89 |



|            |   |    |
|------------|---|----|
| Σχήμα A.1: | Παραδείγματα καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων με χρήση αμφότερων των U-Net και per-pixel MLP latent edge predictors ( $T = 50$ ). Σε κάθε τριάδα παρουσιάζονται από αριστερά προς τα δεξιά: σκίτσο αναφοράς, παραγόμενη εικόνα με χρήση του U-Net και παραγόμενη εικόνα με χρήση του MLP latent edge predictor. . . . . | 95 |
| Σχήμα A.2: | Παραδείγματα καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων με χρήση αμφότερων των U-Net και per-pixel MLP latent edge predictors, ( $T = 50$ ). (συνέχεια) . . . . .   | 96 |
| Σχήμα A.3: | Παραδείγματα καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων με χρήση αμφότερων των U-Net και per-pixel MLP latent edge predictors, ( $T = 50$ ). (συνέχεια) . . . . .   | 97 |
| Σχήμα A.4: | Παραδείγματα καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων με χρήση αμφότερων των U-Net και per-pixel MLP latent edge predictors, ( $T = 50$ ). (συνέχεια) . . . . .   | 98 |
| Σχήμα A.5: | Παραδείγματα καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων με χρήση αμφότερων των U-Net και per-pixel MLP latent edge predictors, ( $T = 50$ ). (συνέχεια) . . . . .   | 99 |



## Κατάλογος Πινάκων

---

|              |   |    |
|--------------|---|----|
| Πίνακας 3.1: | Αρχιτεκτονική attention block του δικτύου U-Net του Σχ.3.2. . . . .   | 43 |
| Πίνακας 4.1: | Αρχιτεκτονική του per-pixel MLP [4]. . . . .  | 50 |
| Πίνακας 4.2: | Αρχιτεκτονική U-Net latent edge predictor. Με $c_o$ , $W$ , $H$ συμβολίζονται το πλήθος των καναλιών, το πλάτος και το ύψος των μητρώων εξόδου του εκάστοτε επιπέδου του δικτύου, αντίστοιχα. Κάθε συνελκτικό στρώμα (Conv $i$ , $i = 1, \dots, 19$ ) ακολουθείται από ένα στρώμα batch normalization και από συνάρτηση ενεργοποίησης ReLU. . . . . | 55 |
| Πίνακας 6.1: | Παράμετροι εκπαίδευσης per-pixel MLP latent edge predictor. . . . .   | 64 |
| Πίνακας 6.2: | Παράμετροι διαδικασίας καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων, με χρήση του per-pixel MLP latent edge predictor. . . . .  | 65 |
| Πίνακας 7.1: | Παράμετροι εκπαίδευσης U-Net latent edge predictor. . . . .   | 74 |
| Πίνακας 7.2: | Παράμετροι διαδικασίας καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεση εικόνων, με χρήση του U-Net latent edge predictor. . . . .   | 75 |
| Πίνακας 8.1: | Αποτελέσματα ποσοτικής αξιολόγησης. . . . .   | 91 |



# Μέρος I

## Εισαγωγή

---



# Κεφάλαιο 1

## Περιγραφή Προβλήματος

---

Το ελεύθερο σχέδιο αποτελεί μέσο έκφρασης της δημιουργικότητας και της φαντασίας των ατόμων, προσφέροντας έναν αφαιρετικό και σχετικά απλό τρόπο αποτύπωσης των διαφορετικών πτυχών του άκρως περίπλοκου και συνεχώς αναπτυσσόμενου σύγχρονου κόσμου. Η δημιουργία σκίτσων χωρίς τη χρήση ειδικών οργάνων σχεδίασης (*freehand sketches*) επιτρέπει στον καθέναν, ανεξαρτήτως καλλιτεχνικών ικανοτήτων, να αποδώσει με έναν προσωπικό και μοναδικό τρόπο, εικόνες από το πεδίο της φαντασίας του. Λόγω της σημασίας του αυτής, το πρόβλημα της μετατροπής αφαιρετικών σκίτσων σε ρεαλιστικές εικόνες παρουσιάζει σημαντικότατο ενδιαφέρον και αποτελεί αντικείμενο μελέτης πολλών ερευνητών. Παρά ταύτα, το εν λόγω πρόβλημα είναι ιδιαίτερος προκλητικό και απαιτητικό, καθώς στη βάση του προσπαθεί να γεφυρώσει το κενό μεταξύ της αφαιρετικής και αδόμητης φύσης των ελεύθερων σκίτσων και του υψηλού επιπέδου λεπτομέρειας των πραγματικών εικόνων. Η γεφύρωση αυτή απαιτεί την ανάπτυξη μεθόδων, ικανών να αντιλαμβάνονται και να εξάγουν χαρακτηριστικά από τις σχετικά περιορισμένες πληροφορίες που παρέχονται από τα σκίτσα.

Στο πλαίσιο αυτό, έχουν προταθεί αρκετές μέθοδοι για την αντιμετώπιση του προβλήματος της καθοδηγούμενης από σκίτσα σύνθεσης εικόνων, οι οποίες βασίζονται στη χρήση GANs [1, 2, 3]. Παρ' όλων των πολλά υποσχόμενων αποτελεσμάτων τους, οι μέθοδοι αυτές είναι αρκετά ελλιπείς, λόγω της απουσίας κειμενικής καθοδήγησης, γεγονός το οποίο οδηγεί σε περιορισμένα και μη ελεγχόμενα ως προς τη νοηματική απόδοση αποτελέσματα. Οι περιορισμοί αυτοί, σε συνδυασμό με το πολύ μεγάλο πλήθος ζευγών εικόνων-σκίττων που απαιτούν κάποιες εξ αυτών για την εκπαίδευση των παραγωγικών μοντέλων, καθιστά επιτακτική την ανάγκη για ανάπτυξη νέων μεθόδων αντιμετώπισης του προβλήματος.

Τα τελευταία χρόνια, η ανάπτυξη των μοντέλων διάχυσης έχει οδηγήσει σε εντυπωσιακά αποτελέσματα σύνθεσης ρεαλιστικών εικόνων υψηλής ευκρίνειας, βάσει αντίστοιχων κειμενικών περιγραφών, με τα μοντέλα αυτά να επιδεικνύουν ξεκάθαρη υπεροχή έναντι των αντίστοιχων GANs. Είναι επομένως αναμενόμενο και φυσικό να μελετηθεί και να εξεταστεί η χρήση τους σε προβλήματα καθοδηγούμενης από σκίτσα σύνθεσης εικόνων.

Στην παρούσα διπλωματική εργασία προτείνεται μία μέθοδος για την καθοδηγούμενη από σκίτσα και κειμενικές περιγραφές σύνθεση εικόνων, η οποία βασίζεται στη χρήση μοντέλων διάχυσης. Στον πυρήνα της προτεινόμενης μεθόδου βρίσκεται το προεκπαιδευμένο

latent μοντέλο *Stable Diffusion*, το οποίο και χρησιμοποιείται για την παραγωγή εικόνων από αντίστοιχες κειμενικές περιγραφές. Για την καθοδήγηση της διαδικασίας σύνθεσης βάσει των περιγραμμάτων των σκίτσων χρησιμοποιείται ένα δίκτυο πρόβλεψης ακμών (*latent edge predictor*). Το εν λόγω δίκτυο λαμβάνει ως είσοδο ένα διάνυμα, το οποίο αποτελείται από τις ενεργοποιήσεις των ενδιάμεσων στρωμάτων του δικτύου αποθρομβοποίησης του *Stable Diffusion* και παρέχει μια εκτίμηση για τις ακμές της παραγόμενης εικόνας στο τρέχον βήμα της αντίστροφης διαδικασίας διάχυσης. Η εκτίμηση αυτή συγκρίνεται εν συνεχεία με το σκίτσο αναφοράς και βάσει της ομοιότητάς τους τροποποιείται καταλλήλως η παραγόμενη εικόνα, ώστε να πλησιάζει κατά το δυνατόν τη χωρική διάταξη του σκίτσου.

Η προαναφερθείσα μέθοδος βασίζεται στους *Voynov et al.* [4], οι οποίοι στην αρχική τους υλοποίηση χρησιμοποιούν ένα per-pixel MLP ως *latent edge predictor*. Για τη βελτίωση της όλης διαδικασίας, στα πλαίσια της εργασίας προτείνεται η υλοποίηση ενός U-Net *latent edge predictor*, ο οποίος είναι ικανός να εξάγει και να λαμβάνει υπόψιν χωρικές συσχετίσεις των pixels εισόδου. Επιπλέον, στο προτεινόμενο πλαίσιο ενσωματώνεται και ένα δίκτυο απλοποίησης σκίτσων (*sketch simplification network*), το οποίο ομαλοποιεί και εξομαλύνει τις ακμές των σκίτσων εισόδου. Με αυτόν τον τρόπο, παρέχεται στο χρήστη η δυνατότητα να προεπεξεργαστεί το σκίτσο πριν την τροφοδότησή του στο μοντέλο διάχυσης. Τα πειραματικά αποτελέσματα, σε συνδυασμό με τις απόψεις που λήφθηκαν από τους χρήστες μέσω ερωτηματολογίων, αποδεικνύουν ότι ο U-Net *latent edge predictor*:

- i) Παράγει πιο ρεαλιστικές και υψηλής ευκρίνειας εικόνες, οι οποίες ακολουθούν πιστά τη χωρική διάταξη των σκίτσων αναφοράς.
- ii) Μειώνει το συνολικό αριθμό βημάτων που απαιτούνται για την παραγωγή ικανοποιητικών αποτελεσμάτων κατά περίπου 80%. Αυτή η βελτίωση οδηγεί σε αναλογική μείωση του συνολικού χρόνου εκτέλεσης.
- iii) Παράγει εικόνες οι οποίες προτιμούνται στην πλειοψηφία τους από την κοινή γνώμη, ως προς το ρεαλισμό, την πιστότητα των ακμών και τη συνολική τους δομή.

## 1.1 Δομή Εργασίας

Η παρούσα διπλωματική διαρθρώνεται σε 9 κεφάλαια, εκ των οποίων το **Κεφάλαιο 1** αποτελεί μια πρώτη εισαγωγή και μια γενική περιγραφή του εξεταζόμενου προβλήματος, παρουσιάζοντας συνοπτικά τα βασικά του μέρη καθώς και την προσέγγιση που ακολουθείται για την αντιμετώπισή του. Τα υπόλοιπα κεφάλαια οργανώνονται σε 2 επιμέρους μέρη, εκ των οποίων το πρώτο καλύπτει το θεωρητικό υπόβαθρο που απαιτείται για την κατανόηση του προβλήματος και της προτεινόμενης μεθόδου, ενώ το δεύτερο περιλαμβάνει την παράθεση των πειραματικών αποτελεσμάτων και των συμπερασμάτων που προκύπτουν από αυτά. Πιο συγκεκριμένα, σε κάθε κεφάλαιο εξετάζονται τα εξής:

- **Κεφάλαιο 2:** Στο κεφάλαιο αυτό παρουσιάζεται το θεωρητικό υπόβαθρο των μοντέλων διάχυσης. Έτσι, αναλύονται οι διαφορετικές κατηγορίες μοντέλων διάχυσης, οι βασικές αρχές λειτουργίας τους και παρουσιάζονται τεχνικές και μέθοδοι επίλυσης της αντίστροφης διαδικασίας διάχυσης.



- **Κεφάλαιο 3:** Στο κεφάλαιο αυτό γίνεται αναφορά στο πρόβλημα της καθοδηγούμενης από κειμενικές περιγραφές σύνθεσης εικόνων και αναλύεται η αρχιτεκτονική και η λειτουργία του μοντέλου Stable Diffusion, το οποίο και αποτελεί βασικό στοιχείο της προτεινόμενης μεθόδου.
- **Κεφάλαιο 4:** Στο κεφάλαιο αυτό περιγράφεται το πρόβλημα τη καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων και αναλύεται η προτεινόμενη μέθοδος αντιμετώπισής του.
- **Κεφάλαιο 5:** Στο κεφάλαιο αυτό περιγράφεται η διαδικασία που ακολουθείται για την κατασκευή του συνόλου των δεδομένων εκπαίδευσης, που χρησιμοποιείται τόσο για την εκπαίδευση του per-pixel MLP όσο και του προτεινόμενου U-Net latent edge predictor.
- **Κεφάλαιο 6:** Στο κεφάλαιο αυτό παρουσιάζονται παραδείγματα καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων, με χρήση του per-pixel MLP latent edge predictor, γίνεται σχολιασμός των αποτελεσμάτων αυτών και εξάγονται συμπεράσματα για την επίδοσή του.
- **Κεφάλαιο 7:** Στο κεφάλαιο αυτό παρουσιάζονται παραδείγματα καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων, με χρήση του U-Net latent edge predictor, γίνεται σχολιασμός των αποτελεσμάτων αυτών και εξάγονται συμπεράσματα για την επίδοσή του.
- **Κεφάλαιο 8:** Στο κεφάλαιο αυτό γίνεται ποιοτική και ποσοτική σύγκριση των αποτελεσμάτων των δύο μοντέλων latent edge predictors, μέσω των οποίων και επιβεβαιώνεται η υπεροχή της προτεινόμενης αρχιτεκτονικής U-Net έναντι της αντίστοιχης αρχιτεκτονικής του MLP.
- **Κεφάλαιο 9:** Στο κεφάλαιο αυτό γίνεται μία συνοπτική ανακεφαλαίωση των όσων αναλύθηκαν και αναπτύχθηκαν στα προηγούμενα κεφάλαια και συνοψίζεται η συμπερασματολογία που προκύπτει από τα ποιοτικά και ποσοτικά αποτελέσματα. Τέλος, γίνεται μία σύντομη αναφορά στη μελλοντική έρευνα και σε τρόπους επέκτασης της λειτουργίας του προτεινόμενου πλαισίου.



Μέρος 

Θεωρητικό Υπόβαθρο

---



# Κεφάλαιο 2

## Diffusion Models

---

Τα μοντέλα διάχυσης (*Diffusion Models*) αποτελούν μία οικογένεια πιθανοτικών παραγωγικών μοντέλων, τα οποία καταστρέφουν προοδευτικά τα δεδομένα εγχύοντας θόρυβο και έπειτα μαθαίνουν να αντιστρέφουν την ανωτέρω διαδικασία, με σκοπό την ανάκτηση των αρχικών δεδομένων. Τα μοντέλα αυτά μπορούν να διακριθούν στις εξής τρεις γενικές κατηγορίες: *denoising diffusion probabilistic models* (DDPMs), *score-based generative models* (SGMs) και *stochastic differential equations* (Score SDEs).

### 2.1 Denoising Diffusion Probabilistic Models

Ένα denoising diffusion probabilistic model (DDPM) [5] κάνει χρήση δύο Μαρκοβιανών αλυσίδων. Η πρώτη εξ αυτών χρησιμοποιείται κατά την προς τα εμπρός διαδικασία διάχυσης και προσθέτει θόρυβο στα αρχικά δεδομένα, ενώ η δεύτερη χρησιμοποιείται κατά την αντίστροφη διαδικασία αποθορυβοποίησης και αναιρεί την προσθήκη του θορύβου, ώστε να ανακτηθούν τα αρχικά δεδομένα. Η πρώτη αλυσίδα σχεδιάζεται ώστε να μετασχηματίζει την κατανομή των διαθέσιμων δεδομένων σε μία πιο απλή κατανομή (π.χ. κανονική κατανομή), ενώ η δεύτερη αντιστρέφει τον ανωτέρω μετασχηματισμό, μαθαίνοντας πυρήνες μετάβασης (*transition kernels*), οι οποίοι παραμετροποιούνται με χρήση βαθέων συνελκτικών νευρωνικών δικτύων.

Τυπικά, δοθείσης μίας κατανομής δεδομένων  $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ , η προς τα εμπρός διαδικασία Μάρκον παράγει μια ακολουθία τυχαίων μεταβλητών  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T$  με πυρήνα μετάβασης  $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ . Χρησιμοποιώντας τον κανόνα της αλυσίδας και τη Μαρκοβιανή ιδιότητα, η από κοινού κατανομή των τυχαίων μεταβλητών  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T$ , δεδομένης της αρχικής τυχαίας μεταβλητής  $\mathbf{x}_0$ , ισούται με,

$$q(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T | \mathbf{x}_0) = \prod_{t=1}^T q(\mathbf{x}_t | \mathbf{x}_{t-1}) \quad (2.1)$$

Ο πυρήνας μετάβασης  $q(\mathbf{x}_t|\mathbf{x}_{t-1})$  της σχέσης (2.1) κατασκευάζεται ώστε να μετασχηματίζει σταδιακά την κατανομή των δεδομένων σε μία εκ των προτέρων γνωστή κατανομή. Μία συνήθης επιλογή για τον πυρήνα μετάβασης, είναι ο Γκαουσιανός πυρήνας, ο οποίος δίνεται από τη σχέση,

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_t = \sqrt{1 - \beta_t}\mathbf{x}_{t-1}, \boldsymbol{\Sigma}_t = \beta_t\mathbf{I}), \quad (2.2)$$

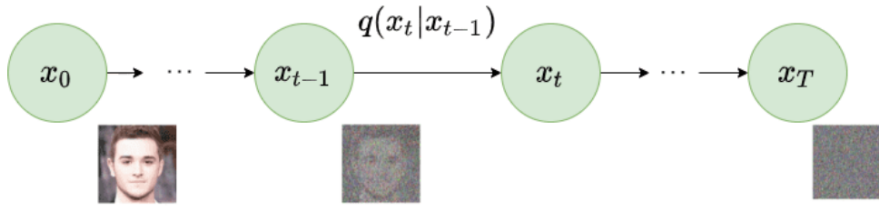
όπου  $\beta_t \in (0, 1)$  μία υπερπαράμετρος, η οποία επιλέγεται κατά το σχεδιασμό του μοντέλου. Η χρήση του Γκαουσιανού πυρήνα μετάβασης, επιτρέπει τον προσδιορισμό μίας αναλυτικής μορφής για την κατανομή  $q(\mathbf{x}_t|\mathbf{x}_0)$  για κάθε  $t \in \{1, \dots, T\}$ . Πιο συγκεκριμένα, θεωρώντας ότι  $\alpha_t := 1 - \beta_t$  και  $\bar{\alpha}_t := \prod_{s=0}^t \alpha_s$ , προκύπτει ότι,

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1 - \bar{\alpha}_t)\mathbf{I}). \quad (2.3)$$

Δοθέντος τώρα ενός αρχικού δείγματος  $\mathbf{x}_0$ , ένα δείγμα  $\mathbf{x}_t$  μπορεί να ληφθεί, δειγματοληπτώντας ένα διάνυσμα Γκαουσιανού θορύβου  $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  και εφαρμόζοντας το μετασχηματισμό

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon. \quad (2.4)$$

Όταν  $\bar{\alpha}_T \approx 0$ , η τυχαία μεταβλητή  $\mathbf{x}_T$  ακολουθεί προσεγγιστικά την κανονική κατανομή, δηλαδή  $q(\mathbf{x}_T) = \int q(\mathbf{x}_T|\mathbf{x}_0)q(\mathbf{x}_0)d\mathbf{x}_0 \approx \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$ . Στο Σχ.2.1 παρουσιάζεται η διαδικασία διάχυσης που αναλύθηκε έως τώρα.



Σχήμα 2.1: Διαδικασία διάχυσης σε DDPM

Διασθητικά, η όλη λειτουργία ενός DDPM βασίζεται στην εξής λογική: κατά την προς τα εμπρός διαδικασία της διάχυσης, γίνεται σταδιακή έγχυση θορύβου στα αρχικά δεδομένα, έως ότου καταστραφεί κάθε δομή, ενώ κατά την αντίστροφη διαδικασία της αποθορυβοποίησης, παράγεται αρχικά ένα διάνυσμα θορύβου από μία εκ των προτέρων γνωστή κατανομή, και εν συνεχεία αφαιρείται σταδιακά αυτό θόρυβος, μέσω μίας εκπαιδευσιμής Μαρκοβιανής αλυσίδας, η οποία κινείται στην αντίθετη χρονική κατεύθυνση από αυτή της αλυσίδας διάχυσης. Πιο συγκεκριμένα, η αντίστροφη Μαρκοβιανή αλυσίδα παραμετροποιείται από μία εκ των προτέρων γνωστή κατανομή  $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$  και από έναν πυρήνα μετάβασης προς μάθηση  $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ . Η επιλογή της κανονικής κατανομής για την παραμετροποίηση της αντίστροφης αλυσίδας, οφείλεται στον τρόπο με τον οποίο κατασκευάζεται η προς τα εμπρός αλυσίδα, η οποία συγκλίνει σε μία κανονική κατανομή  $q(\mathbf{x}_T) \approx \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$ . Ο προς μάθηση πυρήνας μετάβασης  $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ , λαμβάνει τη μορφή,

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)), \quad (2.5)$$

όπου η μεταβλητή  $\theta$  υποδηλώνει τις παραμέτρους του μοντέλου, ενώ η μέση τιμή  $\boldsymbol{\mu}_\theta(\mathbf{x}_t, t)$  και η διασπορά  $\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)$  παραμετροποιούνται μέσω συνελκτικών νευρωνικών δικτύων. Μέσω της αντίστροφης αυτής Μαρκοβιανής αλυσίδας, μπορεί να παραχθεί δείγμα  $\mathbf{x}_0$ , δειγματοληπτώντας αρχικά ένα διάνυσμα θορύβου  $\mathbf{x}_T \sim p(\mathbf{x}_T)$  και δειγματοληπτώντας έπειτα επαναληπτικά από τον πυρήνα μετάβασης  $\mathbf{x}_{t-1} \sim p(\mathbf{x}_{t-1}|\mathbf{x}_t)$ , έως ότου  $t = 1$ .

Προκειμένου ωστόσο η δειγματοληψία αυτή να είναι αποδοτική, θα πρέπει να εκπαιδευτεί η αντίστροφη αλυσίδα, ήτοι να προσδιοριστούν οι παράμετροι  $\theta$ , κατά τρόπο, ώστε η από κοινού κατανομή  $p_\theta(\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_T) := p(\mathbf{x}_T) \prod_{t=1}^T p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$  να προσεγγίζει σε ικανοποιητικό βαθμό την αντίστοιχη από κοινού κατανομή της Μαρκοβιανής αλυσίδας διάχυσης  $q(\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_T) := q(\mathbf{x}_0) \prod_{t=1}^T q(\mathbf{x}_t|\mathbf{x}_{t-1})$ . Αυτό επιτυγχάνεται μέσω της ελαχιστοποίησης του *evidence lower bound* στην αρνητική συνάρτηση λογαριθμικής πιθανοφάνειας:

$$-\log p_\theta(\mathbf{x}_0) = -\log \int p_\theta(\mathbf{x}_{0:T}) d\mathbf{x}_{1:T} \quad (2.6)$$

$$= -\log \int \frac{p_\theta(\mathbf{x}_{0:T})q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} d\mathbf{x}_{1:T} \quad (2.7)$$

$$= -\log \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[ \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \right] \quad (2.8)$$

$$\leq \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[ -\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \right] \quad (\text{Jensen's Inequality}) \quad (2.9)$$

$$= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[ -\log p(\mathbf{x}_T) - \sum_{t \geq 1} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})} \right] \quad (2.10)$$

$$= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[ -\log p(\mathbf{x}_T) - \sum_{t > 1} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})} - \log \frac{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}{q(\mathbf{x}_1|\mathbf{x}_0)} \right] := L \quad (2.11)$$

Σύμφωνα με τον κανόνα του Bayes και την Μαρκοβιανή ιδιότητα, ο πυρήνας μετάβασης  $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ , μπορεί να γραφεί ως,

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0) = \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)q(\mathbf{x}_t|\mathbf{x}_0)}{q(\mathbf{x}_{t-1}|\mathbf{x}_0)} \quad (2.12)$$

Επομένως, χρησιμοποιώντας τη σχέση 2.12, η 2.11 γράφεται ως εξής:

$$\begin{aligned} L &= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[ -\log p(\mathbf{x}_T) - \sum_{t > 1} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} \cdot \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)} - \log \frac{p_\theta(\mathbf{x}_0|\mathbf{x}_1)}{q(\mathbf{x}_1|\mathbf{x}_0)} \right] \\ &= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[ -\log \frac{p(\mathbf{x}_T)}{q(\mathbf{x}_T|\mathbf{x}_0)} - \sum_{t > 1} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)} - \log p_\theta(\mathbf{x}_0|\mathbf{x}_1) \right] \\ &= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[ \underbrace{D_{KL}(q(\mathbf{x}_T|\mathbf{x}_0) \| p(\mathbf{x}_T))}_{L_T} + \sum_{t > 1} \underbrace{D_{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \| p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))}_{L_{t-1}} \right. \\ &\quad \left. \underbrace{-\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)}_{L_0} \right] \end{aligned} \quad (2.13)$$

Συνολικά, η αντικειμενική συνάρτηση μπορεί να γραφεί ως,

$$L = L_T + L_{T-1} + \dots + L_0 \quad (2.14)$$

όπου

$$\begin{aligned} L_T &= D_{KL} (q(\mathbf{x}_T|\mathbf{x}_0) \| p_\theta(\mathbf{x}_T)) \\ L_{t-1} &= D_{KL} (q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \| p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)) \text{ για } 2 \leq t \leq T \\ L_0 &= -\log p_\theta(\mathbf{x}_0|\mathbf{x}_1) \end{aligned}$$

Οι όροι της σχέσης 2.13 ερμηνεύονται ως εξής:

- $L_0$ : πρόκειται για έναν όρο ανακατασκευής, ο οποίος μπορεί να βελτιστοποιηθεί μέσω προσομοίωσης Monte Carlo.
- $L_T$ : πρόκειται για την Kullback-Leibler απόκλιση μεταξύ της κατανομής της τελικής ενθόρυβης εισόδου και της αρχικής πρότερης κανονικής κατανομής. Στην πράξη, ο όρος αυτός είναι σταθερός (τείνει στο μηδέν), καθώς όπως αναφέρθηκε και παραπάνω, η κατανομή της τελικής ενθόρυβης εισόδου προσεγγίζει την κανονική κατανομή.
- $L_{t-1}$ : παράγοντας αποθορυβοποίησης (*denoising matching term*), ο οποίος συμβάλλει στον υπολογισμό του πυρήνα αποθορυβοποίησης  $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ , μέσω της προσέγγισης του ground-truth πυρήνα μετάβασης  $q(\mathbf{x}_t|\mathbf{x}_0)$ . Στην πράξη, ο πυρήνας  $q(\mathbf{x}_t|\mathbf{x}_0)$  καθορίζει τον τρόπο με τον οποίο πρέπει να γίνει η αποθορυβοποίηση μιας θορυβώδους εικόνας  $\mathbf{x}_t$ , δεδομένης της τελικής πλήρους αποθορυβοποιημένης εικόνας  $\mathbf{x}_0$ . Επομένως, ο όρος αυτός ελαχιστοποιείται, όταν τα δύο στάδια αποθορυβοποίησης ταυτίζονται όσο το δυνατόν περισσότερο.

Από τους ανωτέρω όρους, οι  $L_T$  και  $L_0$  παραλείπονται κατά τη διαδικασία της βελτιστοποίησης, καθώς ο πρώτος εξ αυτών είναι σταθερός, ενώ ο δεύτερος όταν δεν λαμβάνεται υπόψιν οδηγεί σε καλύτερα αποτελέσματα, σύμφωνα με τους *Ho et al.* [5] Επομένως, η συνάρτηση απώλειας ισούται με

$$L := L_{t-1} = D_{KL} (q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \| p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)) \quad (2.15)$$

Από τον κανόνα του Bayes ισχύει ότι:

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0)q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)} \quad (2.16)$$

Χρησιμοποιώντας τις σχέσεις 2.2 και 2.3, η δεσμευμένη κατανομή  $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$  γράφεται ισοδυνάμως ως,

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0)q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)} \quad (2.17)$$

$$= \frac{\mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_{t-1}, (1 - \alpha_t)\mathbf{I}) \mathcal{N}(\mathbf{x}_{t-1}; \sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0, (1 - \bar{\alpha}_{t-1})\mathbf{I})}{\mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1 - \bar{\alpha}_t)\mathbf{I})} \quad (2.18)$$



$$\propto \exp \left\{ - \left[ \frac{(\mathbf{x}_t - \sqrt{\alpha_t} \mathbf{x}_{t-1})^2}{2(1-\alpha_t)} + \frac{(\mathbf{x}_{t-1} - \sqrt{\bar{\alpha}_{t-1}} \mathbf{x}_0)^2}{2(1-\bar{\alpha}_{t-1})} - \frac{(\mathbf{x}_t - \sqrt{\bar{\alpha}_t} \mathbf{x}_0)^2}{2(1-\bar{\alpha}_t)} \right] \right\} \quad (2.19)$$

$$\propto \mathcal{N} \left( \mathbf{x}_{t-1}; \underbrace{\frac{\sqrt{\alpha_t} (1 - \bar{\alpha}_{t-1}) \mathbf{x}_t + \sqrt{\bar{\alpha}_{t-1}} (1 - \alpha_t) \mathbf{x}_0}{1 - \bar{\alpha}_t}}_{\mu_q(\mathbf{x}_t, \mathbf{x}_0)}, \underbrace{\frac{(1 - \alpha_t) (1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \mathbf{I}}_{\Sigma_q(t)} \right) \quad (2.20)$$

Από την τελευταία σχέση προκύπτει ότι σε κάθε βήμα, το δείγμα  $\mathbf{x}_{t-1} \sim q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0)$  ακολουθεί κανονική κατανομή με μέση τιμή  $\mu_q(\mathbf{x}_t, \mathbf{x}_0)$  και μητρώο συνδιακύμανσης  $\Sigma_q$ . Προκειμένου να προσεγγιστεί το ground-truth βήμα μετάβασης  $q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0)$ , το βήμα  $p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)$  μπορεί να μοντελοποιηθεί ως μία Γκαουσιανή κατανομή, με μητρώο συνδιακύμανσης  $\Sigma_q(t) = \sigma_q^2(t) \mathbf{I}$ , όπου

$$\sigma_q^2(t) = \frac{(1 - \alpha_t) (1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t}. \quad (2.21)$$

Στο σημείο αυτό υπενθυμίζεται ότι η απόσταση Kulback-Leibler μεταξύ δύο Γκαουσιανών κατανομών δίνεται από την κάτωθι σχέση:

$$\begin{aligned} D_{\text{KL}} \left( \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) \parallel \mathcal{N}(\mathbf{y}; \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y) \right) &= \\ &= \frac{1}{2} \left[ \log \frac{|\boldsymbol{\Sigma}_y|}{|\boldsymbol{\Sigma}_x|} - d + \text{tr} \left( \boldsymbol{\Sigma}_y^{-1} \boldsymbol{\Sigma}_x \right) + \left( \boldsymbol{\mu}_y - \boldsymbol{\mu}_x \right)^T \boldsymbol{\Sigma}_y^{-1} \left( \boldsymbol{\mu}_y - \boldsymbol{\mu}_x \right) \right] \end{aligned} \quad (2.22)$$

Στην περίπτωσή μας, εφόσον μπορεί να εξασφαλιστεί ότι τα μητρώα συνδιακύμανσης των δύο Γκαουσιανών θα είναι ίσα, η ελαχιστοποίηση της απόκλισης KL μεταξύ των δύο κατανομών, έγκειται στην ελαχιστοποίηση της διαφοράς των μέσων τιμών τους. Επομένως, ο όρος  $L_{t-1}$  μπορεί να γραφεί ως,

$$L_{t-1} = \mathbb{E}_q \left[ \frac{1}{2\sigma_q^2(t)} \left[ \|\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_q\|_2^2 \right] \right], \quad (2.23)$$

όπου οι όροι  $\boldsymbol{\mu}_q$  και  $\boldsymbol{\mu}_\theta$ , αποτελούν συντομογραφίες των όρων  $\boldsymbol{\mu}_q(\mathbf{x}_t, \mathbf{x}_0)$  και  $\boldsymbol{\mu}_\theta(\mathbf{x}_t, t)$  αντίστοιχα. Από τη σχέση 2.20 είναι γνωστό ότι η μέση τιμή  $\boldsymbol{\mu}_q(\mathbf{x}_t, \mathbf{x}_0)$  ισούται με,

$$\boldsymbol{\mu}_q(\mathbf{x}_t, \mathbf{x}_0) = \frac{\sqrt{\alpha_t} (1 - \bar{\alpha}_{t-1}) \mathbf{x}_t + \sqrt{\bar{\alpha}_{t-1}} (1 - \alpha_t) \mathbf{x}_0}{1 - \bar{\alpha}_t}. \quad (2.24)$$

Μετασχηματίζοντας τη σχέση 2.4 λαμβάνεται η κάτωθι αναπαράσταση για το  $\mathbf{x}_0$ :

$$\mathbf{x}_0 = \frac{\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}}{\sqrt{\bar{\alpha}_t}}. \quad (2.25)$$

Αντικαθιστώντας τώρα την ανωτέρω σχέση στη σχέση 2.24 προκύπτει ότι,

$$\boldsymbol{\mu}_q(\mathbf{x}_t, \mathbf{x}_0) = \frac{\sqrt{\bar{\alpha}_t} (1 - \bar{\alpha}_{t-1}) \mathbf{x}_t + \sqrt{\bar{\alpha}_{t-1}} (1 - \alpha_t) \mathbf{x}_0}{1 - \bar{\alpha}_t} \quad (2.26)$$

$$= \frac{\sqrt{\bar{\alpha}_t} (1 - \bar{\alpha}_{t-1}) \mathbf{x}_t + \sqrt{\bar{\alpha}_{t-1}} (1 - \alpha_t) \frac{\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}}{\sqrt{\bar{\alpha}_t}}}{1 - \bar{\alpha}_t} \quad (2.27)$$

$$= \frac{\sqrt{\bar{\alpha}_t} (1 - \bar{\alpha}_{t-1}) \mathbf{x}_t + (1 - \alpha_t) \frac{\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}}{\sqrt{\bar{\alpha}_t}}}{1 - \bar{\alpha}_t} \quad (2.28)$$

$$= \frac{\sqrt{\bar{\alpha}_t} (1 - \bar{\alpha}_{t-1}) \mathbf{x}_t}{1 - \bar{\alpha}_t} + \frac{(1 - \alpha_t) \mathbf{x}_t}{(1 - \bar{\alpha}_t) \sqrt{\bar{\alpha}_t}} - \frac{(1 - \alpha_t) \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}}{(1 - \bar{\alpha}_t) \sqrt{\bar{\alpha}_t}} \quad (2.29)$$

$$= \left( \frac{\sqrt{\bar{\alpha}_t} (1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} + \frac{1 - \alpha_t}{(1 - \bar{\alpha}_t) \sqrt{\bar{\alpha}_t}} \right) \mathbf{x}_t - \frac{(1 - \alpha_t) \sqrt{1 - \bar{\alpha}_t}}{(1 - \bar{\alpha}_t) \sqrt{\bar{\alpha}_t}} \boldsymbol{\epsilon} \quad (2.30)$$

$$= \left( \frac{\alpha_t (1 - \bar{\alpha}_{t-1})}{(1 - \bar{\alpha}_t) \sqrt{\bar{\alpha}_t}} + \frac{1 - \alpha_t}{(1 - \bar{\alpha}_t) \sqrt{\bar{\alpha}_t}} \right) \mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t} \sqrt{\bar{\alpha}_t}} \boldsymbol{\epsilon} \quad (2.31)$$

$$= \frac{\alpha_t - \bar{\alpha}_t + 1 - \alpha_t}{(1 - \bar{\alpha}_t) \sqrt{\bar{\alpha}_t}} \mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t} \sqrt{\bar{\alpha}_t}} \boldsymbol{\epsilon} \quad (2.32)$$

$$= \frac{1 - \bar{\alpha}_t}{(1 - \bar{\alpha}_t) \sqrt{\bar{\alpha}_t}} \mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t} \sqrt{\bar{\alpha}_t}} \boldsymbol{\epsilon} \quad (2.33)$$

$$= \frac{1}{\sqrt{\bar{\alpha}_t}} \mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t} \sqrt{\bar{\alpha}_t}} \boldsymbol{\epsilon} \quad (2.34)$$

Επομένως, η μέση τιμή του πυρήνα αποθρομβοποίησης  $\boldsymbol{\mu}_\theta(\mathbf{x}_t, t)$  μπορεί να τεθεί ίση με,

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\bar{\alpha}_t}} \mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t} \sqrt{\bar{\alpha}_t}} \hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, t). \quad (2.35)$$

Συνδυάζοντας τις σχέσεις 2.23, 2.34 και 2.35, ο όρος  $L_{t-1}$  γράφεται ισοδυνάμως ως,

$$L_{t-1} = \mathbb{E}_q \left[ \frac{1}{2\sigma_q^2(t)} \left[ \|\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_q\|_2^2 \right] \right] \quad (2.36)$$

$$= \mathbb{E}_q \left[ \left\| \frac{1}{\sqrt{\bar{\alpha}_t}} \mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t} \sqrt{\bar{\alpha}_t}} \hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, t) - \frac{1}{\sqrt{\bar{\alpha}_t}} \mathbf{x}_t + \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t} \sqrt{\bar{\alpha}_t}} \boldsymbol{\epsilon} \right\|_2^2 \right] \quad (2.37)$$

$$= \mathbb{E}_q \left[ \frac{1}{2\sigma_q^2(t)} \left\| \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t} \sqrt{\bar{\alpha}_t}} \boldsymbol{\epsilon} - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t} \sqrt{\bar{\alpha}_t}} \hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, t) \right\|_2^2 \right] \quad (2.38)$$

$$= \mathbb{E}_q \left[ \frac{1}{2\sigma_q^2(t)} \left\| \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t} \sqrt{\alpha_t}} (\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, t)) \right\|_2^2 \right] \quad (2.39)$$

$$= \mathbb{E}_q \left[ \frac{1}{2\sigma_q^2(t)} \frac{(1 - \alpha_t)^2}{(1 - \bar{\alpha}_t) \alpha_t} \|\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, t)\|_2^2 \right] \quad (2.40)$$

$$= \mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon}} \left[ \lambda(t) \left\| \boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}_\theta \left( \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}, t \right) \right\|_2^2 \right] \quad (2.41)$$

όπου

$$\lambda(t) = \frac{1}{2\sigma_q^2(t)} \frac{(1 - \alpha_t)^2}{(1 - \bar{\alpha}_t) \alpha_t}. \quad (2.42)$$

Στη σχέση 2.41, ο όρος  $\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, t)$  αναφέρεται σε ένα νευρωνικό δίκτυο, το οποίο μαθαίνει να προβλέπει το διάνυσμα θορύβου  $\boldsymbol{\epsilon}$ , βάσει του οποίου προκύπτει το ενθόρυβο δείγμα  $\mathbf{x}_t$  από το αρχικό δείγμα  $\mathbf{x}_0$ . Κατά αυτό τον τρόπο προκύπτει ότι η εκπαίδευση του μοντέλου έγκειται στην πρόβλεψη του θορύβου που χρησιμοποιήθηκε για την παραγωγή του ενθόρυβου δείγματος. Στο σημείο αυτό αναφέρεται ότι οι *Ho et al.* [5] διαπίστωσαν εμπειρικά ότι η εκπαίδευση του μοντέλου διάχυσης είναι πιο αποδοτική, αν χρησιμοποιηθεί μια απλοποιημένη εκδοχή της συνάρτησης της σχέσης 2.41, στην οποία παραλείπονται οι όροι στάθμισης,

$$L_{t-1}^{simple} = \mathbb{E}_{t \sim \mathcal{U}[1, T], \mathbf{x}_0 \sim q(\mathbf{x}_0), \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[ \left\| \boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}_\theta \left( \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}, t \right) \right\|_2^2 \right]. \quad (2.43)$$

Επομένως, η τελική συνάρτηση απώλειας για την εκπαίδευση του μοντέλου διάχυσης θα είναι η,

$$L_{simple} = L_{t-1}^{simple} + C, \quad (2.44)$$

όπου η παράμετρος  $C$  είναι μία σταθερά, η οποία και αγνοείται κατά τη διάρκεια της βελτιστοποίησης.

---

#### Αλγόριθμος 2.1: Training

---

- 1: **repeat**
  - 2:  $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
  - 3:  $t \sim \text{Uniform}(\{1, \dots, T\})$
  - 4:  $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  - 5: Take gradient descent step on  $\nabla_\theta \left\| \boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}_\theta \left( \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}, t \right) \right\|_2^2$
  - 6: **until** converged
- 

---

#### Αλγόριθμος 2.2: Sampling

---

- 1:  $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  - 2: **for**  $t = T, \dots, 1$  **do**
  - 3:  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  if  $t > 1$ , else  $\mathbf{z} = \mathbf{0}$
  - 4:  $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left( \mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, t) \right) + \sigma_q(t) \mathbf{z}$
  - 5: **end for**
  - 6: **return**  $\mathbf{x}_0$
-

## 2.2 Score-based Generative Models

### 2.2.1 Score Matching

Στον πυρήνα των score-based generative μοντέλων [6, 7, 8] βρίσκεται η έννοια του *Score Matching* [9]. Η μέθοδος αυτή ανήκει στην κατηγορία των non-likelihood-based μεθόδων και χρησιμοποιείται για τη δειγματοληψία από άγνωστες κατανομές, προσπαθώντας να άρει τους περιορισμούς που εισάγουν οι αντίστοιχες likelihood-based μέθοδοι. Αυτό επιτυγχάνεται μέσω της εκμάθησης της *score function* της άγνωστης συνάρτησης πυκνότητας πιθανότητας.

#### Ορισμός: Score Function

Η score function μίας κατανομής δεδομένων  $p(\mathbf{x})$  δίνεται από τη σχέση,

$$f(\mathbf{x}) = \nabla_{\mathbf{x}} \log p(\mathbf{x}).$$

Η συνάρτηση αυτή πρόκειται ουσιαστικά για ένα διανυσματικό πεδίο, το οποίο δείχνει προς τις κατευθύνσεις εκείνες, κατά τις οποίες η συνάρτηση πυκνότητας πιθανότητας παρουσιάζει το μεγαλύτερο ρυθμό αύξησης. Προκειμένου να γίνει αντιληπτή η σημασία της βελτιστοποίησης της συνάρτησης score, θα αξιοποιηθεί η θεωρία των energy-based μοντέλων [10]. Στα μοντέλα αυτά, η κατανομή πιθανότητας  $p_{\theta}(\mathbf{x})$  μπορεί να γραφεί ως,

$$p_{\theta}(\mathbf{x}) = \frac{1}{Z_{\theta}} e^{-f_{\theta}(\mathbf{x})}, \quad (2.45)$$

όπου η συνάρτηση  $f_{\theta}(\mathbf{x})$  είναι μία παραμετροποιήσιμη συνάρτηση (συνάρτηση ενέργειας), η οποία μοντελοποιείται μέσω νευρωνικών δικτύων και  $Z_{\theta}$  μία παράμετρος κανονικοποίησης, η οποία εξασφαλίζει ότι  $\int p_{\theta}(\mathbf{x}) d\mathbf{x} = 1$ . Ένας τρόπος να προσεγγιστεί μία τέτοια κατανομή είναι μέσω τεχνικών μεγιστοποίησης της πιθανοφάνειας, κάτι το οποίο προϋποθέτει ωστόσο τον υπολογισμό της σταθεράς κανονικοποίησης  $Z_{\theta} = \int e^{-f_{\theta}(\mathbf{x})} d\mathbf{x}$ , ο οποίος μπορεί να μην είναι εφικτός για περίπλοκες συναρτήσεις  $f_{\theta}(\mathbf{x})$ .

Ένας τρόπος με τον οποίο μπορεί να αποφευχθεί ο υπολογισμός ή η μοντελοποίηση της σταθεράς κανονικοποίησης, είναι μέσω της χρήσης ενός νευρωνικού δικτύου  $s_{\theta}(\mathbf{x})$ , το οποίο καλείται να μάθει τη score function της κατανομής των δεδομένων. Η χρήση του δικτύου αυτού πηγάζει από την παρατήρηση, ότι αν λάβουμε την παράγωγο και στα δύο μέλη της εξίσωσης 2.45 προκύπτει η ακόλουθη σχέση:

$$\nabla_{\mathbf{x}} \log p_{\theta}(\mathbf{x}) = \nabla_{\mathbf{x}} \log \left( \frac{1}{Z_{\theta}} e^{-f_{\theta}(\mathbf{x})} \right) \quad (2.46)$$

$$= \nabla_{\mathbf{x}} \log \frac{1}{Z_{\theta}} + \nabla_{\mathbf{x}} \log e^{-f_{\theta}(\mathbf{x})} \quad (2.47)$$

$$= -\nabla_{\mathbf{x}} f_{\theta}(\mathbf{x}) \quad (2.48)$$

$$\approx s_{\theta}(\mathbf{x}) \quad (2.49)$$

Το μοντέλο  $s_\theta(\mathbf{x})$  της ανωτέρω σχέσης μπορεί να βελτιστοποιηθεί ελαχιστοποιώντας την απόκλιση Fisher από την ground truth score function, η οποία ισούται με:

$$J(\theta) := \frac{1}{2} \mathbb{E}_{p(\mathbf{x})} [\|s_\theta(\mathbf{x}) - \nabla_{\mathbf{x}} \log p(\mathbf{x})\|_2^2]. \quad (2.50)$$

Παρά ταύτα, το κύριο πρόβλημα που παρουσιάζεται σε αυτή την περίπτωση είναι ότι η score function  $\nabla_{\mathbf{x}} \log p(\mathbf{x})$  δεν είναι γνωστή, καθώς εξαρτάται άμεσα από την άγνωστη κατανομή των δεδομένων  $p(\mathbf{x})$ . Στο πλαίσιο αυτό, οι Hyvärinen et al. [9] έδειξαν ότι η σχέση 2.50 μπορεί να γραφεί ισοδυνάμως ως,

$$J(\theta) = \frac{1}{2} \mathbb{E}_{p(\mathbf{x})} \left[ \text{tr}(\nabla_{\mathbf{x}} s_\theta(\mathbf{x})) + \frac{1}{2} \|s_\theta(\mathbf{x})\|_2^2 \right]. \quad (2.51)$$

Η παραπάνω σχέση μπορεί να προσδιοριστεί χρησιμοποιώντας μεθόδους Monte Carlo, δειγματοληπτώντας από την κατανομή  $p(\mathbf{x})$ , αφού αυτή εξαρτάται μόνο από το  $s_\theta(\mathbf{x})$ .

### 2.2.2 Sliced Score Matching

Ο προσδιορισμός του όρου του ίχνους  $\text{tr}(\nabla_{\mathbf{x}} s_\theta(\mathbf{x}))$  στη σχέση 2.51 είναι πολύ δύσκολος υπολογιστικά, ειδικά σε περιπτώσεις όπου τα δεδομένα  $\mathbf{x}$  είναι υψηλών διαστάσεων. Προκειμένου να αντιμετωπιστεί η δυσκολία αυτή, χρησιμοποιείται μία εναλλακτική μέθοδος score matching, η οποία καλείται *Sliced Score Matching* [11].

Από τούδε και στο εξής, η score function της κατανομής των δεδομένων θα συμβολίζεται με  $s_d(\mathbf{x}) = \nabla_{\mathbf{x}} \log p(\mathbf{x})$ . Η βασική ιδέα της μεθόδου sliced score matching έγκειται στην προβολή των διανυσματικών πεδίων  $s_\theta(\mathbf{x})$  και  $s_d(\mathbf{x})$  σε κάποια τυχαία κατεύθυνση  $\mathbf{v}$  και στη μετέπειτα σύγκριση της μέσης διαφοράς των προβολών αυτών. Πιο συγκεκριμένα, ως εναλλακτική της απόκλισης Fisher της σχέσης 2.50, χρησιμοποιείται η ακόλουθη αντικειμενική συνάρτηση:

$$L(\theta; p_{\mathbf{v}}) := \frac{1}{2} \mathbb{E}_{p_{\mathbf{v}}, \mathbb{E}_{p(\mathbf{x})}} \left[ \left( \mathbf{v}^T s_\theta(\mathbf{x}) - \mathbf{v}^T s_d(\mathbf{x}) \right)^2 \right], \quad (2.52)$$

όπου  $\mathbf{v} \sim p_{\mathbf{v}}$  και  $\mathbf{x} \sim p_d$  ανεξάρτητα. Επίσης απαιτούμε  $\mathbb{E}_{p_{\mathbf{v}}}[\mathbf{v}\mathbf{v}^T] > 0$  και  $\mathbb{E}_{p_{\mathbf{v}}}[\|\mathbf{v}\|_2^2] < \infty$ . Η σχέση 2.52 γράφεται ισοδυνάμως ως,

$$L(\theta; p_{\mathbf{v}}) = \frac{1}{2} \mathbb{E}_{p_{\mathbf{v}}, \mathbb{E}_{p(\mathbf{x})}} \left[ \left( \mathbf{v}^T s_\theta(\mathbf{x}) - \mathbf{v}^T s_d(\mathbf{x}) \right)^2 \right] \quad (2.53)$$

$$= \frac{1}{2} \mathbb{E}_{p_{\mathbf{v}}, \mathbb{E}_{p(\mathbf{x})}} \left[ \left( \mathbf{v}^T s_\theta(\mathbf{x}) \right)^2 + \left( \mathbf{v}^T s_d(\mathbf{x}) \right)^2 - 2 \left( \mathbf{v}^T s_\theta(\mathbf{x}) \right) \left( \mathbf{v}^T s_d(\mathbf{x}) \right) \right] \quad (2.54)$$

$$= \mathbb{E}_{p_{\mathbf{v}}, \mathbb{E}_{p(\mathbf{x})}} \left[ \frac{1}{2} \left( \mathbf{v}^T s_\theta(\mathbf{x}) \right)^2 - \left( \mathbf{v}^T s_\theta(\mathbf{x}) \right) \left( \mathbf{v}^T s_d(\mathbf{x}) \right) \right] + \mathbf{C}, \quad (2.55)$$

όπου ο όρος  $s_d(\mathbf{x})$  απορροφάται στην σταθερά  $\mathbf{C}$ , καθώς δεν εξαρτάται από την παράμετρο  $\theta$ .

Παρατηρείται τώρα ότι:

$$- \mathbb{E}_{p_{\mathbf{v}}} \mathbb{E}_{p(\mathbf{x})} \left[ \left( \mathbf{v}^T \mathbf{s}_{\theta}(\mathbf{x}) \right) \left( \mathbf{v}^T \mathbf{s}_d(\mathbf{x}) \right) \right] \quad (2.56)$$

$$= - \mathbb{E}_{p_{\mathbf{v}}} \int \left[ \left( \mathbf{v}^T \mathbf{s}_{\theta}(\mathbf{x}) \right) \left( \mathbf{v}^T \mathbf{s}_d(\mathbf{x}) \right) p(\mathbf{x}) d\mathbf{x} \right] \quad (2.57)$$

$$= - \mathbb{E}_{p_{\mathbf{v}}} \left[ \int \left( \mathbf{v}^T \mathbf{s}_{\theta}(\mathbf{x}) \right) \left( \mathbf{v}^T \nabla_{\mathbf{x}} \log p(\mathbf{x}) \right) p(\mathbf{x}) d\mathbf{x} \right] \quad (2.58)$$

$$= - \mathbb{E}_{p_{\mathbf{v}}} \left[ \int \left( \mathbf{v}^T \mathbf{s}_{\theta}(\mathbf{x}) \right) \left( \mathbf{v}^T \nabla_{\mathbf{x}} p(\mathbf{x}) \right) d\mathbf{x} \right] \quad (2.59)$$

$$= - \mathbb{E}_{p_{\mathbf{v}}} \left[ \int \left( \mathbf{v}^T \mathbf{s}_{\theta}(\mathbf{x}) \right) \left( \mathbf{v}^T \nabla_{\mathbf{x}} p(\mathbf{x}) \right) d\mathbf{x} \right] \quad (2.60)$$

$$= - \mathbb{E}_{p_{\mathbf{v}}} \left[ \sum_i \int \left( \mathbf{v}^T \mathbf{s}_{\theta}(\mathbf{x}) \right) \left( v_i \frac{\partial p(\mathbf{x})}{\partial x_i} \right) d\mathbf{x} \right] \quad (2.61)$$

$$= \mathbb{E}_{p_{\mathbf{v}}} \left[ \int \mathbf{v}^T \mathbf{s}_{\theta}(\mathbf{x}) \mathbf{v} \cdot p(\mathbf{x}) d\mathbf{x} \right] \quad (2.62)$$

$$= \mathbb{E}_{p_{\mathbf{v}}} \mathbb{E}_{p(\mathbf{x})} \left[ \mathbf{v}^T \mathbf{s}_{\theta}(\mathbf{x}) \mathbf{v} \right], \quad (2.63)$$

όπου η σχέση 2.62 προκύπτει έπειτα από εφαρμογή ολοκλήρωσης κατά μέλη. Κατά αυτό τον τρόπο καταλήγουμε στην ισοδύναμη αντικειμενική συνάρτηση,

$$J(\boldsymbol{\theta}; p_{\mathbf{v}}) := \mathbb{E}_{p_{\mathbf{v}}} \mathbb{E}_{p(\mathbf{x})} \left[ \mathbf{v}^T \nabla_{\mathbf{x}} \mathbf{s}_{\theta}(\mathbf{x}) \mathbf{v} + \frac{1}{2} \|\mathbf{s}_{\theta}(\mathbf{x})\|_2^2 \right], \quad (2.64)$$

για την οποία αποδεικνύεται [11], ότι κάτω από ορισμένες προϋποθέσεις κανονικότητας, ισχύει

$$L(\boldsymbol{\theta}; p_{\mathbf{v}}) = J(\boldsymbol{\theta}; p_{\mathbf{v}}) + C, \quad (2.65)$$

όπου  $C$  μία παράμετρος σταθερή ως προς  $\boldsymbol{\theta}$ . Η νέα αυτή αντικειμενική συνάρτηση δεν παρουσιάζει κάποια εξάρτηση από την άγνωστη παράμετρο  $s_d(\mathbf{x})$ . Ως εκ τούτου, μπορεί να θεωρηθεί ο ακόλουθος αμερόληπτος εκτιμητής της:

$$\hat{J}_{N,M}(\boldsymbol{\theta}; p_{\mathbf{v}}) = \frac{1}{N} \frac{1}{M} \sum_{i=1}^N \sum_{j=1}^M \left[ \mathbf{v}_{ij}^T \nabla_{\mathbf{x}} \mathbf{s}_{\theta}(\mathbf{x}_i) \mathbf{v}_{ij} + \frac{1}{2} \|\mathbf{s}_{\theta}(\mathbf{x}_i)\|_2^2 \right], \quad (2.66)$$

όπου  $\mathbf{v}_{ij}$  η  $j$ -οστή προβολή του δείγματος  $\mathbf{x}_i$  από την κατανομή  $p_{\mathbf{v}}$ , με  $1 \leq j \leq M$  και  $1 \leq i \leq N$ . Στα πλαίσια της ανάλυσης, η κατανομή  $p_{\mathbf{v}}$  αντιστοιχεί είτε σε μία πολυμεταβλητή κανονική κατανομή είτε σε μία πολυμεταβλητή κατανομή Rademacher.

### 2.2.3 Denoising Score Matching

Η μέθοδος του *Denoising Score Matching* [12] αποσκοπεί και αυτή στην αποφυγή του υπολογισμού του όρου  $\text{tr}(\nabla_{\mathbf{x}} s_{\theta}(\mathbf{x}))$  και βασίζεται στην προσθήκη θορύβου στα δεδομένα. Πιο συγκεκριμένα, βάσει της μεθόδου αυτής, προστίθεται αρχικά θόρυβος σε ένα δείγμα  $\mathbf{x}$  βάσει μιας προκαθορισμένης κατανομής θορύβου  $q_{\sigma}(\tilde{\mathbf{x}}|\mathbf{x})$  και έπειτα προσεγγίζεται η κατανομή του ενθόρυβου πλέον δείγματος  $q_{\sigma}(\tilde{\mathbf{x}}) := \int q_{\sigma}(\tilde{\mathbf{x}}|\mathbf{x})p(\mathbf{x})d\mathbf{x}$ . Η αντικειμενική συνάρτηση στην περίπτωση αυτή ισούται με:

$$J(\theta; q_{\sigma}) := \frac{1}{2} \mathbb{E}_{q_{\sigma}(\tilde{\mathbf{x}}|\mathbf{x})p(\mathbf{x})} [\|s_{\theta}(\tilde{\mathbf{x}}) - \nabla_{\tilde{\mathbf{x}}} \log q_{\sigma}(\tilde{\mathbf{x}}|\mathbf{x})\|_2^2]. \quad (2.67)$$

### 2.2.4 Langevin Dynamics

Αφού εκπαιδευτεί κάποιο δίκτυο  $s_{\theta}(\mathbf{x})$  για τον προσδιορισμό της *score function* της κατανομής των διαθέσιμων δεδομένων, δειγματοληψία από την κατανομή αυτή μπορεί να υλοποιηθεί μέσω της μεθόδου *Langevin Dynamics*. Η μέθοδος αυτή πρόκειται για μία Markov Chain Monte Carlo μέθοδο και χρησιμοποιείται για τη δειγματοληψία από μία στάσιμη κατανομή, για την οποία μπορούμε να λάβουμε αποδοτικά παραγώγους ως προς την πιθανότητα των δειγμάτων  $\mathbf{x}$ .

Στη μέθοδο αυτή, γίνεται εκκίνηση από ένα αρχικό σημείο  $\mathbf{x}_0 \sim \pi(\mathbf{x})$ , το οποίο δειγματοληπτείται από μία πρότερη κατανομή  $\pi$  και στη συνέχεια λαμβάνονται επαναληπτικά νέα σημεία σύμφωνα με τη σχέση,

$$\tilde{\mathbf{x}}_t = \tilde{\mathbf{x}}_{t-1} + \frac{\alpha}{2} \nabla_{\mathbf{x}} \log p(\tilde{\mathbf{x}}_{t-1}) + \sqrt{\alpha} \mathbf{z}_t, \text{ για } t = 1, \dots, T \quad (2.68)$$

όπου  $\mathbf{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ . Η προσθήκη θορύβου στη διαδικασία της δειγματοληψίας είναι απαραίτητη, καθώς χωρίς αυτή δεν εξασφαλίζεται η σύγκλιση σε κάποια στάσιμη κατανομή. Αποδεικνύεται ότι καθώς  $\alpha \rightarrow 0$  και  $T \rightarrow \infty$ , η κατανομή της διαδικασίας συγκλίνει στην πραγματική κατανομή των δεδομένων  $p(\mathbf{x})$  [13].

### 2.2.5 Noise Conditional Score Networks (NCSN)

Έχοντας ολοκληρώσει στο σημείο αυτό την ανάλυση των μεθόδων score matching και της δειγματοληψίας μέσω Langevin dynamics, θα εξεταστεί πως οι τεχνικές αυτές αξιοποιούνται από μία ειδική κατηγορία παραγωγικών μοντέλων, αυτή των *Noise Conditional Score Networks (NCSN)* [6].

Ακολουθώντας τους συμβολισμούς της ενότητας 2.1, θεωρείται ότι τα διαθέσιμα δεδομένα ακολουθούν την κατανομή  $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ . Έστω ακόμη  $\{\sigma_i\}_{i=1}^L$  μία θετική γεωμετρική ακολουθία με  $L$  επίπεδα θορύβου, η οποία ικανοποιεί τη σχέση  $\frac{\sigma_1}{\sigma_2} = \dots = \frac{\sigma_{L-1}}{\sigma_L} > 1$ . Καθένα εξ αυτών των επιπέδων αντιστοιχεί σε Γακουσιανό θόρυβο, ο οποίος προστίθεται στα διαθέσιμα δεδομένα, με κατανομή  $q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \mathbf{x}_0, \sigma_t^2 \mathbf{I})$ . Κατά αυτό τον τρόπο δημιουργείται μία ακολουθία θορυβωδών δεδομένων με συναρτήσεις πυκνότητας πιθανότητας  $q(\mathbf{x}_1), q(\mathbf{x}_2), \dots, q(\mathbf{x}_T)$ , όπου  $q(\mathbf{x}_t) := \int q(\mathbf{x}_t)q(\mathbf{x}_0)d\mathbf{x}_0$ . Ένα noise-conditional

score network πρόκειται για ένα βαθύ νευρωνικό δίκτυο  $s_\theta(\mathbf{x}_t, \sigma_t)$ , το οποίο εκπαιδεύεται ώστε να υπολογίζει τη συνάρτηση  $\nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t)$ . Στην περίπτωση αυτή και βάσει της σχέσης 2.67, η αντικειμενική συνάρτηση για κάθε επίπεδο θορύβου  $\sigma_t$  λαμβάνει τη μορφή,

$$\ell(\theta; \sigma_t) := \frac{1}{2} \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0)} \mathbb{E}_{\mathbf{x}_t \sim q(\mathbf{x}_t | \mathbf{x}_0)} \left[ \left\| s_\theta(\mathbf{x}_t, \sigma_t) + \frac{\mathbf{x}_t - \mathbf{x}_0}{\sigma_t^2} \right\|_2^2 \right], \quad (2.69)$$

ενώ η ολική αντικειμενική συνάρτηση κατά μήκος όλων των επιπέδων θορύβου τη μορφή,

$$L(\theta; \{\sigma_i\}_{i=1}^L) := \frac{1}{L} \sum_{i=1}^L \lambda(\sigma_i) \ell(\theta; \sigma_i). \quad (2.70)$$

Αφού εκπαιδευτεί το noise-conditional score network, μπορεί να χρησιμοποιηθεί η μέθοδος Langevin dynamics προκειμένου να ληφθούν καινούρια δείγματα. Παρά ταύτα, αυτή η μέθοδος δειγματοληψίας δεν είναι αποτελεσματική σε περιπτώσεις όπου υπάρχουν περιοχές χαμηλής πυκνότητας στην κατανομή των δεδομένων, με αποτέλεσμα να μην εξασφαλίζεται η σύγκλιση στην πραγματική κατανομή. Προς την κατεύθυνση αυτή, οι Song και Ermon [6] προτείνουν μία εναλλακτική προσέγγιση δειγματοληψίας, η οποία βασίζεται στην προσομοιωμένη ανόπτηση [14, 15] και ονομάζεται **annealed Langevin dynamics**. Όπως φαίνεται και στον αλγόριθμο 2.3, η διαδικασία ξεκινά με την αρχικοποίηση των δειγμάτων από μία πρότερη κατανομή (π.χ. ομοιόμορφος θόρυβος). Στη συνέχεια εκτελείται ο αλγόριθμος Langevin dynamics για τη λήψη δειγμάτων από την κατανομή  $q(\mathbf{x}_1)$  με βήμα  $\alpha_1$ . Έπειτα, ο αλγόριθμος Langevin dynamics εκτελείται και πάλι για δειγματοληψία από την κατανομή  $q(\mathbf{x}_2)$ , θεωρώντας ως αρχικά δείγματα, τα τελευταία δείγματα που λήφθηκαν από την κατανομή  $q(\mathbf{x}_1)$ . Η διαδικασία αυτή επαναλαμβάνεται έως ότου η δειγματοληψία οδηγηθεί στην κατανομή  $q(\mathbf{x}_L)$ , η οποία στο όριο όπου  $\sigma_L \approx 0$  προσεγγίζει την πραγματική κατανομή των δεδομένων. Σε κάθε βήμα, η παράμετρος  $\alpha_i$  ανανεώνεται σύμφωνα με τη σχέση,  $\alpha_i = \epsilon \cdot \sigma_i^2 / \sigma_L^2$ .

---

### Αλγόριθμος 2.3: Annealed Langevin dynamics.

---

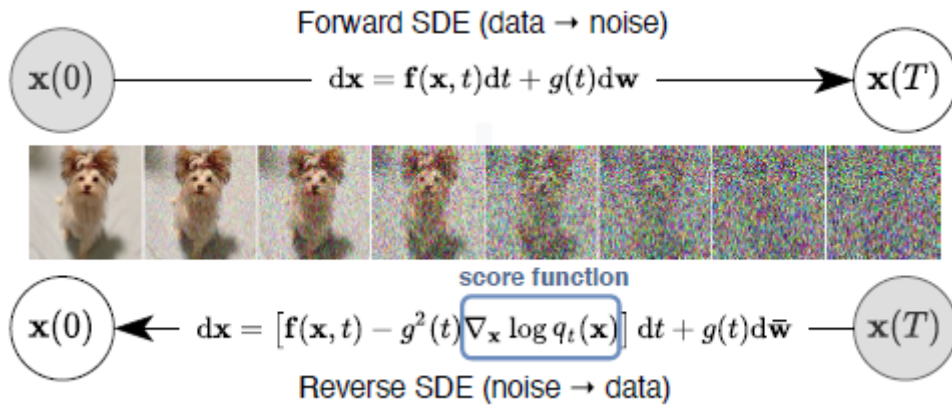
**Require:**  $\{\sigma_i\}_{i=1}^L, \epsilon, T$

- 1: Initialize  $\tilde{\mathbf{x}}_0$
  - 2: **for**  $i \leftarrow 1$  to  $L$  **do**
  - 3:      $\alpha_i = \epsilon \cdot \sigma_i^2 / \sigma_L^2$
  - 4:     **for**  $t \leftarrow 1$  to  $T$  **do**
  - 5:         Draw  $\mathbf{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  - 6:          $\tilde{\mathbf{x}}_t \leftarrow \tilde{\mathbf{x}}_{t-1} + \frac{\alpha_i}{2} s_\theta(\tilde{\mathbf{x}}_{t-1}, \sigma_i) + \sqrt{\alpha_i} \mathbf{z}_t$
  - 7:     **end for**
  - 8:      $\tilde{\mathbf{x}}_0 \leftarrow \tilde{\mathbf{x}}_T$
  - 9: **end for**
  - 10: **return**  $\tilde{\mathbf{x}}_T$
-



## 2.3 Score-based Generative Modeling μέσω Στοχαστικών Διαφορικών Εξισώσεων

Έως το σημείο αυτό, μελετήθηκε η κατηγορία των score-based generative μοντέλων, τα οποία προσθέτουν θόρυβο στα διαθέσιμα δεδομένα μέσω διαδοχικών, πεπερασμένων επιπέδων και εν συνεχεία προσπαθούν να προσεγγίσουν τη score function της κατανομής  $q(\mathbf{x}_t)$ . Προκειμένου να χρησιμοποιηθούν καινούριες μέθοδοι δειγματοληψίας και προκειμένου να επεκταθούν συνολικά οι ικανότητες των score-based generative μοντέλων, η προσθήκη θορύβου στα διαθέσιμα δεδομένα μπορεί να πραγματοποιηθεί σε θεωρητικά άπειρα χρονικά επίπεδα. Στην περίπτωση αυτή, οι διαδικασίες θορυβοποίησης και αποθορυβοποίησης αποτελούν λύσεις στοχαστικών διαφορικών εξισώσεων (SDEs) [8]. Πιο συγκεκριμένα, η διαδικασία αποθορυβοποίησης ικανοποιεί μία reverse-time SDE, η οποία μπορεί να προκύψει από την SDE της πρόσθιας διαδικασίας θορυβοποίησης, δεδομένου του score των περιθώριων κατανομών πιθανότητας, το οποίο εκφράζεται ως συνάρτηση του χρόνου. Είναι εφικτό επομένως η reverse-time SDE να προσεγγιστεί εκπαιδεύοντας ένα χρονοεξαρτώμενο (*time-dependent*) νευρωνικό δίκτυο για τον υπολογισμό των scores. Εν συνεχεία η παραγωγή των δειγμάτων υλοποιείται λαμβάνοντας λύσεις της διαφορικής αυτής εξίσωσης. Η εν λόγω διαδικασία παρουσιάζεται στο Σχ.2.2.



**Σχήμα 2.2:** Λειτουργία Score-based Generative μοντέλων μέσω στοχαστικών διαφορικών εξισώσεων.

Πριν προχωρήσουμε σε περαιτέρω ανάλυση του τρόπου λειτουργίας των score-based generative μοντέλων της κατηγορίας αυτής, κρίνεται σκόπιμο να γίνει μια αναφορά στη θεωρία των στοχαστικών διαφορικών εξισώσεων. Μία στοχαστική διαφορική εξίσωση πρόκειται για μία εξίσωση της μορφής,

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w}. \quad (2.71)$$

Στην παραπάνω σχέση (2.71), η παράμετρος  $\mathbf{f}$  ονομάζεται συντελεστής ολίσθησης (*drift coefficient*) και μοντελοποιεί το ντετερμινιστικό μέρος της SDE, καθορίζοντας το βαθμό στον οποίο η διαδικασία  $d\mathbf{x}$  αναμένεται να μεταβληθεί στο χρόνο. Αντιστοίχως η παράμετρος  $g(t)$  ονομάζεται συντελεστής διάχυσης (*diffusion coefficient*) και αναπαριστά το

στοχαστικό μέρος της SDE, καθορίζοντας το εύρος της διαδικασίας θορυβοποίησης σε βάθος χρόνου. Τέλος, η παράμετρος  $d\mathbf{w}$  αναφέρεται σε μία κίνηση *Brown* και ως εκ τούτου, η ποσότητα  $g(t)d\mathbf{w}$  αναπαριστά τη συνολική διαδικασία θορυβοποίησης.

Στα πλαίσια των score-based generative μοντέλων, η διαδικασία διάχυσης  $\{\mathbf{x}_t\}_{t=0}^T$  πρέπει να κατασκευαστεί με τέτοιο τρόπο, έτσι ώστε  $\mathbf{x}_0 \sim q(\mathbf{x}_0)$  η κατανομή των αρχικών δεδομένων και  $\mathbf{x}_T \sim q(\mathbf{x}_T)$  η Γκαουσιανή κατανομή του θορύβου, η οποία και είναι ανεξάρτητη της αρχική κατανομής των δεδομένων. Έπειτα και εφόσον κάθε SDE παρουσιάζει και μία αντίστοιχη αντίστροφη SDE, δείγματα από την αρχική κατανομή  $q(\mathbf{x}_0)$  μπορεί να ανακτηθούν ξεκινώντας από την τελική κατανομή θορύβου και διατρέχοντας την αντίστροφη SDE, η οποία και δίνεται από τη σχέση [16],

$$d\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g(t)^2 \nabla_{\mathbf{x}} \log q_t(\mathbf{x})] dt + g(t)d\bar{\mathbf{w}}, \quad (2.72)$$

όπου  $\bar{\mathbf{w}}$  μία κίνηση Brown, η οποία ρέει προς την αντίθετη χρονική κατεύθυνση, από  $T$  προς  $0$  και  $dt$  ένα απειροστό αρνητικό χρονικό βήμα. Όταν το score κάθε περιθώριας κατανομής  $\nabla_{\mathbf{x}} \log q_t(\mathbf{x})$ , είναι γνωστό για κάθε  $t$ , η αντίστροφη SDE μπορεί να προσδιοριστεί μέσω της εξίσωσης 2.72 και εν συνεχεία να προσομοιωθεί για τη λήψη δειγμάτων από την αρχική κατανομή  $q(\mathbf{x}_0)$ .

Προκειμένου τώρα να προσεγγιστεί αυτή η συνάρτηση score, μπορεί να εκπαιδευτεί ένα χρονοεξαρτώμενο score-based μοντέλο  $s_{\theta,t}$ , χρησιμοποιώντας ως αντικειμενική συνάρτηση την,

$$\arg \min_{\theta} \mathbb{E}_{t \sim \mathcal{U}[0,T]} \left\{ \lambda(t) \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0)} \mathbb{E}_{\mathbf{x}_t \sim q(\mathbf{x}_t | \mathbf{x}_0)} \left[ \left\| s_{\theta}(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log q_{0t}(\mathbf{x}_t | \mathbf{x}_0) \right\|_2^2 \right] \right\}, \quad (2.73)$$

όπου  $\lambda : [0, T] \rightarrow \mathbb{R}_{>0}$  μία θετική συνάρτηση,  $t$  μία παράμετρος η οποία δειγματοληπτείται από μία ομοιόμορφη κατανομή στο εύρος  $[0, T]$ ,  $\mathbf{x}_0 \sim q(\mathbf{x}_0)$  και  $\mathbf{x}_t \sim q_{0t}(\mathbf{x}_t | \mathbf{x}_0)$ .

### 2.3.1 SMLD ως Διακριτοποίηση Variance Exploding (VE) SDE

Στην περίπτωση των score-matching μοντέλων με Langevin dynamics, όταν χρησιμοποιείται ένα σύνολο  $N$  επιπέδων θορύβου, κάθε πυρήνας  $q(\mathbf{x}_i | \mathbf{x}_0)$  μπορεί να προκύψει από μία Μαρκοβιανή αλυσίδα, η οποία περιγράφεται από την ακόλουθη σχέση:

$$\mathbf{x}_i = \mathbf{x}_{i-1} + \sqrt{\sigma_i^2 - \sigma_{i-1}^2} \mathbf{z}_{i-1}, \quad i = 1, \dots, N, \quad (2.74)$$

όπου  $\mathbf{z}_{i-1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ,  $\mathbf{x}_0 \sim q(\mathbf{x}_0)$  και  $\sigma_0 = 0$ . Στο όριο όπου  $N \rightarrow \infty$ , η Μαρκοβιανή αλυσίδα  $\{\mathbf{x}_i\}_{i=1}^N$  μετατρέπεται σε συνεχή στοχαστική διαδικασία  $\{\mathbf{x}(t)\}_{t=0}^1$ , και οι τυπικές αποκλίσεις των επιπέδων θορύβου  $\{\sigma_i\}_{i=1}^N$  σε συνάρτηση του χρόνου  $\sigma(t)$ , ενώ ο θόρυβος  $\mathbf{z}_i$  μετατρέπεται σε  $\mathbf{z}(t)$ , όπου  $t$  μία μεταβλητή συνεχούς χρόνου, με  $t \in [0, 1]$ . Αν θεωρηθεί ότι  $\mathbf{x}(\frac{i}{N}) = \mathbf{x}_i$ ,  $\sigma(\frac{i}{N}) = \sigma_i$  και  $\mathbf{z}(\frac{i}{N}) = \mathbf{z}_i$  για  $i = 1, \dots, N$ , η εξίσωση 2.74 γράφεται ισοδυνάμως ως,

$$\mathbf{x}(t + \Delta t) = \mathbf{x}(t) + \sqrt{\sigma^2(t + \Delta t) - \sigma^2(t)} \mathbf{z}(t) \approx \mathbf{x}(t) + \sqrt{\frac{d[\sigma^2(t)]}{dt} \Delta t} \mathbf{z}(t), \quad (2.75)$$

όπου  $\Delta t = \frac{1}{N}$  και  $t \in \{0, \frac{1}{N}, \dots, \frac{N-1}{N}\}$ . Η προσεγγιστική ισότητα της ανωτέρω σχέσης ισχύει για  $\Delta \ll 1$ . Στο όριο όπου  $\Delta t \rightarrow 0$ , η εξίσωση συγκλίνει στην,

$$dx = \sqrt{\frac{d[\sigma^2(t)]}{dt}} dw, \quad (2.76)$$

η οποία και πρόκειται για μία *Variance Exploding* στοχαστική διαφορική εξίσωση.

Επομένως, η διαδικασία της προσθήκης θορύβου στην περίπτωση των μοντέλων SMLD μπορεί να περιγραφεί από τη διακριτοποιημένη εκδοχή των Variance Exploding στοχαστικών διαφορικών εξισώσεων.

### 2.3.2 DDPM ως Διακριτοποίηση Variance Preserving (VP) SDE

Στην περίπτωση των DDPMs τώρα, οι πυρήνες θορύβου  $q(x_i|x_0)$  μπορούν να περιγραφούν από μία διακριτή Μαρκοβιανή αλυσίδα, η οποία εκφράζεται από τη σχέση,

$$x_i = \sqrt{1 - \beta_i} x_{i-1} + \sqrt{\beta_i} z_{i-1}, \quad i = 1, \dots, N, \quad (2.77)$$

όπου  $z_{i-1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ . Προκειμένου να ληφθεί το όριο καθώς  $N \rightarrow \infty$ , ορίζονται τα βοηθητικά επίπεδα θορύβου  $\{\bar{\beta}_i = N\beta_i\}_{i=1}^N$ , οπότε και η εξίσωση 2.77 γράφεται ισοδυνάμως,

$$x_i = \sqrt{1 - \frac{\bar{\beta}_i}{N}} x_{i-1} + \sqrt{\frac{\bar{\beta}_i}{N}} z_{i-1}, \quad i = 1, \dots, N. \quad (2.78)$$

Στο όριο καθώς  $N \rightarrow \infty$ , οι παράμετροι  $\{\bar{\beta}_i\}_{i=1}^N$  γίνονται συνάρτηση του χρόνου  $\beta(t)$ , όπου  $t$  συνεχής μεταβλητή στο διάστημα  $t \in [0, 1]$ . Αν θεωρηθεί ότι  $x(\frac{i}{N}) = x_i$ ,  $\beta(\frac{i}{N}) = \bar{\beta}_i$  και  $z(\frac{i}{N}) = z_i$  για  $i = 1, \dots, N$ , η Μαρκοβιανή αλυσίδα της σχέσης 2.78 μπορεί να διατυπωθεί εκ νέου ως εξής:

$$\begin{aligned} x(t + \Delta t) &= \sqrt{1 - \beta(t + \Delta t)\Delta t} x(t) + \sqrt{\beta(t + \Delta t)\Delta t} z(t) \\ &\approx x(t) - \frac{1}{2}\beta(t + \Delta t)\Delta t x(t) + \sqrt{\beta(t + \Delta t)\Delta t} z(t) \\ &\approx x(t) - \frac{1}{2}\beta(t)\Delta t x(t) + \sqrt{\beta(t)\Delta t} z(t), \end{aligned} \quad (2.79)$$

όπου όπως και πριν  $\Delta t = \frac{1}{N}$  και  $t \in \{0, \frac{1}{N}, \dots, \frac{N-1}{N}\}$ . Η προσεγγιστική ισότητα της ανωτέρω σχέσης ισχύει για  $\Delta \ll 1$ . Στο όριο όπου  $\Delta t \rightarrow 0$ , η εξίσωση 2.79 συγκλίνει στην,

$$dx = -\frac{1}{2}\beta(t)x dt + \sqrt{\beta(t)} dw. \quad (2.80)$$

Επομένως, η διαδικασία της προσθήκης θορύβου στην περίπτωση των DDPMs μπορεί να περιγραφεί από τη διακριτοποιημένη εκδοχή των Variance Preserving στοχαστικών διαφορικών εξισώσεων.

### 2.3.3 Επίλυση της Αντίστροφης SDE

Έπειτα από την εκπαίδευση του χρονοεξαρτώμενου score-based μοντέλου  $s_{\theta,t}$  (ενότητα 2.3), αυτό μπορεί να αξιοποιηθεί για την κατασκευή της αντίστροφης SDE της σχέσης 2.72 και έπειτα, χρησιμοποιώντας κατάλληλες αριθμητικές μεθόδους, να γίνει λήψη δειγμάτων από την κατανομή  $q(\mathbf{x}_0)$ .

#### Αριθμητικοί Επιλυτές Στοχαστικών Διαφορικών Εξισώσεων

Οι αριθμητικοί επιλυτές (*numerical solvers*) παρέχουν προσεγγιστικές τροχιές από στοχαστικές διαφορικές εξισώσεις. Υπάρχουν αρκετές αριθμητικές μέθοδοι γενικού σκοπού για την επίλυση στοχαστικών διαφορικών εξισώσεων, όπως η μέθοδος *Euler-Maruyama* και οι στοχαστικές μέθοδοι *Runge-Kutta*. Οποιαδήποτε από τις μεθόδους αυτές μπορεί να χρησιμοποιηθεί για τη λήψη δειγμάτων από την αντίστροφη SDE.

Η μέθοδος δειγματοληψίας η οποία χρησιμοποιείται στην περίπτωση των DDPMs (Αλγόριθμος 2.2), αποτελεί ουσιαστικά μια ειδική περίπτωση διακριτοποίησης της αντίστροφης VP SDE της σχέσης 2.80. Παρά ταύτα, η εξαγωγή παρόμοιων κανόνων για τη δειγματοληψία από νέες SDEs, δεν είναι πάντοτε τετριμμένη. Προς την κατεύθυνση αυτή, οι *Song et al.* [8] προτείνουν μία ειδική μέθοδο για δειγματοληψία, τους *reverse diffusion samplers*, οι οποίοι διακριτοποιούν την αντίστροφη SDE με αντίστοιχο τρόπο με αυτό της πρόσθιας SDE. Πιο συγκεκριμένα, δοθείσης μιας SDE

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + \mathbf{G}(t)d\mathbf{w}, \quad (2.81)$$

γίνεται η υπόθεση ότι ο ακόλουθος επαναληπτικός κανόνας αποτελεί μία διακριτοποίησή της:

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \mathbf{f}_i(\mathbf{x}_i) + \mathbf{G}_i \mathbf{z}_i, \quad i = 0, 1, \dots, N-1, \quad (2.82)$$

όπου  $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ . Βάσει τώρα της εξίσωσης 2.82, για τη διακριτοποίηση της αντίστροφης SDE,

$$d\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - \mathbf{G}(t)\mathbf{G}(t)^T \nabla_{\mathbf{x}} \log q_t(\mathbf{x})] dt + \mathbf{G}(t)d\bar{\mathbf{w}}, \quad (2.83)$$

μπορεί να χρησιμοποιηθεί ο ακόλουθος επαναληπτικός κανόνας

$$\mathbf{x}_i = \mathbf{x}_{i+1} - \mathbf{f}_{i+1}(\mathbf{x}_{i+1}) + \mathbf{G}_{i+1} \mathbf{G}_{i+1}^T \mathbf{s}_{\theta}(\mathbf{x}_{i+1}, i+1) + \mathbf{G}_{i+1} \mathbf{z}_{i+1}, \quad i = 0, 1, \dots, N-1, \quad (2.84)$$

όπου το εκπαιδευμένο score-based μοντέλο  $\mathbf{s}_{\theta}(\mathbf{x}_i, i)$  εξαρτάται κάθε φορά από τον αριθμό της επανάληψης  $i$ .

Εφαρμόζοντας τώρα την σχέση 2.84 στις σχέσεις 2.77 και 2.74, λαμβάνεται ένα νέο σύνολο αριθμητικών επιλυτών για τις αντίστροφες VP και VE SDEs.

### Predictor-Corrector Samplers

Σε αντίθεση με τη γενική περίπτωση των στοχαστικών διαφορικών εξισώσεων, στην περίπτωση μας διαθέτουμε επιπλέον πληροφορία, η οποία μπορεί να χρησιμοποιηθεί για τη βελτίωση της ποιότητας των λύσεων που λαμβάνονται. Εφόσον υπάρχει διαθέσιμο ένα score-based μοντέλο  $s_\theta(\mathbf{x}, t) \approx \nabla_{\mathbf{x}} \log q_t(\mathbf{x})$ , μπορούν να αξιοποιηθούν score-based MCMC μέθοδοι, προκειμένου να ληφθούν δείγματα απευθείας από την κατανομή  $q_t$  και να διορθωθεί η λύση των αριθμητικών επιλυτών. Στο πλαίσιο αυτό, σε κάθε χρονικό βήμα, ο αριθμητικός επιλυτής παρέχει αρχικά μία εκτίμηση του δείγματος του επόμενου χρονικού βήματος, λειτουργώντας ως "predictor". Έπειτα, η score-based MCMC προσέγγιση διορθώνει την περιθώρια κατανομή του εκτιμηθέντος δείγματος, λειτουργώντας ως "corrector". Όσον αφορά τώρα στις τεχνικές λεπτομέρειες, ως predictor μπορεί να χρησιμοποιηθεί οποιοσδήποτε αριθμητικός επιλυτής, ενώ ως corrector οποιαδήποτε score-based MCMC μέθοδος. Για παράδειγμα, όταν χρησιμοποιείται ο reverse diffusion sampler ως predictor και η μέθοδος annealed Langevin dynamics ως corrector, προκύπτουν οι αλγόριθμοι 2.4 και 2.5 για τη δειγματοληψία από VE και VP SDEs αντίστοιχα.

---

#### Αλγόριθμος 2.4: PC sampling (VE SDE)

---

```

1:  $\mathbf{x}_N \sim \mathcal{N}(\mathbf{0}, \sigma_N^2 \mathbf{I})$ 
2: for  $i = N - 1$  to 0 do
3:    $\mathbf{x}'_i \leftarrow \mathbf{x}_{i+1} + (\sigma_{i+1}^2 - \sigma_i^2) \mathbf{s}_\theta(\mathbf{x}_{i+1}, \sigma_{i+1})$ 
4:    $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
5:    $\mathbf{x}_i \leftarrow \mathbf{x}'_i + \sqrt{\sigma_{i+1}^2 - \sigma_i^2} \mathbf{z}$  Predictor
6:   for  $j = 1$  to  $M$  do Corrector
7:      $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
8:      $\mathbf{x}_i \leftarrow \mathbf{x}_i + \frac{\alpha_j}{2} \mathbf{s}_\theta(\mathbf{x}_i, \sigma_i) + \sqrt{\alpha_j} \mathbf{z}$ 
9:   end for
10: end for
11: return  $\mathbf{x}_0$ 

```

---



---

#### Αλγόριθμος 2.5: PC sampling (VP SDE)

---

```

1:  $\mathbf{x}_N \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
2: for  $i = N - 1$  to 0 do
3:    $\mathbf{x}'_i \leftarrow \left(2 - \sqrt{1 - \beta_{i+1}}\right) \mathbf{x}_{i+1} + \beta_{i+1} \mathbf{s}_\theta(\mathbf{x}_{i+1}, i + 1)$ 
4:    $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
5:    $\mathbf{x}_i \leftarrow \mathbf{x}'_i + \sqrt{\beta_{i+1}} \mathbf{z}$  Predictor
6:   for  $j = 1$  to  $M$  do Corrector
7:      $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
8:      $\mathbf{x}_i \leftarrow \mathbf{x}_i + \frac{\alpha_j}{2} \mathbf{s}_\theta(\mathbf{x}_i, \sigma_i) + \sqrt{\alpha_j} \mathbf{z}$ 
9:   end for
10: end for
11: return  $\mathbf{x}_0$ 

```

---

Στο σημείο αυτό και για λόγους πληρότητας, θα αναλυθούν οι τρόποι με τους οποίους προκύπτουν οι διακριτοποιήσεις των αντίστροφων διαφορικών των predictors στις δύο περιπτώσεις των αλγορίθμων 2.4 και 2.5.

Στην περίπτωση των VE SDEs, συγκρίνοντας τις εξισώσεις 2.82 και 2.74 προκύπτει άμεσα ότι  $f_{i+1}(\mathbf{x}_{i+1}) = 0$  και  $G_{i+1} = \sqrt{\sigma_{i+1}^2 - \sigma_i^2}$  και ως εκ τούτου η σχέση 2.84 γράφεται ισοδυνάμως ως,

$$\mathbf{x}_i^{VE} = \mathbf{x}_{i+1} + \left(\sigma_{i+1}^2 - \sigma_i^2\right) s_{\theta}(\mathbf{x}_{i+1}, \sigma_{i+1}) + \sqrt{\sigma_{i+1}^2 - \sigma_i^2} \mathbf{z}. \quad (2.85)$$

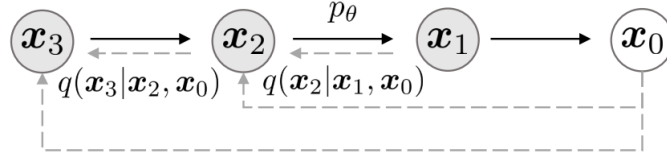
Στην περίπτωση των VP SDEs τώρα, συγκρίνοντας αντιστοίχως τις εξισώσεις 2.82 και 2.77 προκύπτει ότι  $f_{i+1}(\mathbf{x}_{i+1}) = \left(\sqrt{1 - \beta_{i+1}} - 1\right) \mathbf{x}_{i+1}$  και  $G_{i+1} = \sqrt{\beta_{i+1}}$  και ως εκ τούτου η σχέση 2.84 γράφεται ισοδυνάμως ως,

$$\begin{aligned} \mathbf{x}_i &= \mathbf{x}_{i+1} - \left(\sqrt{1 - \beta_{i+1}} - 1\right) \mathbf{x}_{i+1} + \left(\sqrt{\beta_{i+1}}\right)^2 s_{\theta}(\mathbf{x}_{i+1}, \sigma_{i+1}) + \sqrt{\beta_{i+1}} \mathbf{z} \\ &= \left(2 - \sqrt{1 - \beta_{i+1}}\right) \mathbf{x}_{i+1} + \beta_{i+1} s_{\theta}(\mathbf{x}_{i+1}, i + 1) + \sqrt{\beta_{i+1}} \mathbf{z}. \end{aligned} \quad (2.86)$$

## 2.4 Denoising Diffusion Implicit Models

Στην ενότητα αυτή θα γίνει αναφορά σε μία ειδική κατηγορία μοντέλων διάχυσης, τα λεγόμενα *Denoising Diffusion Implicit Models (DDIMs)* [17]. Έως τώρα, παρουσιάστηκαν και αναλύθηκαν τα DDPMs, καθώς και τα NCSNs, τα οποία προσπαθούν αμφοτέρω να παράξουν ρεαλιστικά δείγματα βάσει της κατανομής των αρχικών δεδομένων. Βασικό μειονέκτημα των μοντέλων αυτών είναι, ότι προκειμένου να παράξουν τέτοια δείγματα υψηλής ποιότητας, απαιτείται ένας πολύ μεγάλος αριθμός επαναλήψεων. Για παράδειγμα, στην περίπτωση των DDPMs, η διαδικασία παραγωγής νέων δειγμάτων, ακολουθεί την αντίστροφη πορεία από αυτή της διαδικασίας διάχυσης, η οποία ωστόσο μπορεί να αποτελείται από χιλιάδες επίπεδα θορύβου. Η επανάληψη και η αντίστροφη διέλευση από όλα τα επιμέρους αυτά επίπεδα είναι πολύ πιο αργή συγκριτικά με την περίπτωση των GANs, στα οποία μία και μόνο διέλευση μέσα από το δίκτυο αρκεί για την παραγωγή ενός ρεαλιστικού δείγματος υψηλής ποιότητας. Για την αντιμετώπιση του προβλήματος αυτού και τη μείωση του χρόνου που απαιτείται για την παραγωγή νέων δειγμάτων, οι *Song et al.* προτείνουν μία νέα κατηγορία μοντέλων, αυτή των DDIMs, τα οποία αποτελούν μία επέκταση των DDPMs, υπό την έννοια ότι σε αυτά, η διαδικασία διάχυσης είναι πλέον **μη Μαρκοβιανή**.

Βασική παρατήρηση για την κατασκευή των μοντέλων αυτών είναι ότι η αντικειμενική συνάρτηση των DDPMs της σχέσης 2.43 εξαρτάται μόνο από τις περιθώριες κατανομές  $q(\mathbf{x}_t | \mathbf{x}_0)$  και όχι άμεσα από την από κοινού κατανομή  $q(\mathbf{x}_{1:T} | \mathbf{x}_0)$ . Εφόσον λοιπόν υπάρχουν διαφορετικές από κοινού κατανομές με τις ίδιες περιθώριες κατανομές, μπορούν να εξερευνηθούν διαφορετικές forward διαδικασίες, οι οποίες είναι μη Μαρκοβιανές και οι οποίες οδηγούν σε νέες αντίστροφες διαδικασίες παραγωγής, όπως φαίνεται στο Σχ.2.3.



Σχήμα 2.3: Μη Μαρκοβιανές διαδικασίες

Ας θεωρήσουμε τώρα την ακόλουθη οικογένεια  $Q$  κατανομών, οι οποίες δεικτοδοτούνται από ένα πραγματικό διάνυσμα  $\sigma \in \mathbb{R}_{\geq 0}^T$ :

$$q_\sigma(\mathbf{x}_{1:T}|\mathbf{x}_0) := q_\sigma(\mathbf{x}_T|\mathbf{x}_0) \prod_{t=2}^T q_\sigma(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0), \quad (2.87)$$

όπου  $q_\sigma(\mathbf{x}_T|\mathbf{x}_0) = \mathcal{N}(\sqrt{\alpha_T}\mathbf{x}_0, (1 - \alpha_T)\mathbf{I})$  για κάθε  $t > 1$  και

$$q_\sigma(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}\left(\sqrt{\alpha_{t-1}}\mathbf{x}_0 + \sqrt{1 - \alpha_{t-1} - \sigma_t^2} \cdot \frac{\mathbf{x}_t - \sqrt{\alpha_t}\mathbf{x}_0}{\sqrt{1 - \alpha_t}}, \sigma_t^2\mathbf{I}\right). \quad (2.88)$$

Η μέση τιμή στην παραπάνω σχέση επιλέγεται, ώστε να εξασφαλίζεται ότι  $q_\sigma(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\sqrt{\alpha_t}\mathbf{x}_0, (1 - \alpha_t)\mathbf{I})$  για κάθε  $t$ . Στην περίπτωση αυτή, η forward διαδικασία, βάσει του κανόνα του Bayes, περιγράφεται από τη σχέση,

$$\begin{aligned} q_\sigma(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0) &= \frac{q_\sigma(\mathbf{x}_{t-1}, \mathbf{x}_0|\mathbf{x}_t)q_\sigma(\mathbf{x}_t)}{q_\sigma(\mathbf{x}_{t-1}, \mathbf{x}_0)} \\ &= \frac{q_\sigma(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_0) \frac{1}{q_\sigma(\mathbf{x}_t)} q_\sigma(\mathbf{x}_t)}{q_\sigma(\mathbf{x}_{t-1}|\mathbf{x}_0)q_\sigma(\mathbf{x}_0)} \\ &= \frac{q_\sigma(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)q_\sigma(\mathbf{x}_t|\mathbf{x}_0)q_\sigma(\mathbf{x}_0)}{q_\sigma(\mathbf{x}_{t-1}|\mathbf{x}_0)q_\sigma(\mathbf{x}_0)} \\ &= \frac{q_\sigma(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)q_\sigma(\mathbf{x}_t|\mathbf{x}_0)}{q_\sigma(\mathbf{x}_{t-1}|\mathbf{x}_0)}. \end{aligned} \quad (2.89)$$

Σε αντίθεση με τη διαδικασία διάχυσης της σχέσης 2.1, η forward διαδικασία στην περίπτωση της σχέσης 2.87 **δεν είναι πλέον Μαρκοβιανή**, καθώς κάθε δείγμα  $\mathbf{x}_t$  μπορεί να εξαρτάται τόσο από το  $\mathbf{x}_{t-1}$  όσο και από το  $\mathbf{x}_0$ . Η τιμή της παραμέτρου  $\sigma$  καθορίζει τη στοχαστικότητα της διαδικασίας. Στην οριακή περίπτωση όπου  $\sigma \rightarrow \mathbf{0}$ , η διαδικασία είναι ντετερμινιστική και εφόσον τα  $\mathbf{x}_0$  και  $\mathbf{x}_t$  παρατηρηθούν για κάποιο  $t$ , το  $\mathbf{x}_{t-1}$  καθίσταται γνωστό και σταθερό.

Στη συνέχεια καθορίζεται μία εκπαιδεύσιμη διαδικασία παραγωγής  $p_\theta(\mathbf{x}_{0:T})$ . Διαισθητικά, δοθέντος ενός θορυβώδους δείγματος  $\mathbf{x}_t$ , γίνεται αρχικά μία πρόβλεψη για το αντίστοιχο  $\mathbf{x}_0$  και έπειτα η πρόβλεψη αυτή χρησιμοποιείται για τη λήψη του δείγματος  $\mathbf{x}_{t-1}$ , μέσω της αντίστροφης δεσμευμένης κατανομής  $q_\sigma(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ , η οποία και έχει ήδη οριστεί.

Για κάποιο  $\mathbf{x}_0 \sim q(\mathbf{x}_0)$  και  $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ , το θορυβώδες δείγμα  $\mathbf{x}_t$  μπορεί να εκτιμηθεί μέσω της σχέσης 2.4, όπως φαίνεται παρακάτω:

$$f_{\theta}(\mathbf{x}_t) := (\mathbf{x}_t - \sqrt{1 - \alpha_t} \cdot \hat{\boldsymbol{\epsilon}}_{\theta}(\mathbf{x}_t)) / \sqrt{\alpha_t}. \quad (2.90)$$

Δοθείσης επομένως μίας σταθερής πρότερης κατανομής  $p_{\theta} = \mathcal{N}(\mathbf{0}, \mathbf{I})$ , μπορεί να οριστεί η ακόλουθη διαδικασία παραγωγής,:

$$p_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t) = \begin{cases} \mathcal{N}(f_{\theta}(\mathbf{x}_t), \sigma_t^2 \mathbf{I}), & \text{αν } t = 1 \\ q_{\sigma}(\mathbf{x}_{t-1} | \mathbf{x}_t, f_{\theta}(\mathbf{x}_t)), & \text{αλλιώς,} \end{cases} \quad (2.91)$$

όπου η κατανομή  $q_{\sigma}(\mathbf{x}_{t-1} | \mathbf{x}_t, f_{\theta}(\mathbf{x}_t))$  ορίζεται σύμφωνα με τη σχέση 2.88, στην οποία το δείγμα  $\mathbf{x}_0$  αντικαθίσταται από την εκτίμησή του  $f_{\theta}(\mathbf{x}_t)$ .

Βάσει τώρα της σχέσης 2.91, η παραγωγή ενός δείγματος  $\mathbf{x}_{t-1}$ , δεδομένου του δείγματος  $\mathbf{x}_t$ , γίνεται σύμφωνα με τη σχέση,

$$\mathbf{x}_{t-1} = \underbrace{\sqrt{\alpha_{t-1}} \left( \frac{\mathbf{x}_t - \sqrt{1 - \alpha_t} \hat{\boldsymbol{\epsilon}}_{\theta}(\mathbf{x}_t)}{\sqrt{\alpha_t}} \right)}_{\text{"πρόβλεψη } \mathbf{x}_0 \text{"}} + \underbrace{\sqrt{1 - \alpha_{t-1} - \sigma_t^2} \cdot \hat{\boldsymbol{\epsilon}}_{\theta}(\mathbf{x}_t)}_{\text{"κατεύθυνση προ το } \mathbf{x}_t \text{"}} + \underbrace{\sigma_t \boldsymbol{\epsilon}}_{\text{τυχαίος θόρυβος}}, \quad (2.92)$$

όπου  $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ . Διαφορετικές τιμές της παραμέτρου  $\sigma$  οδηγούν σε διαφορετικές διαδικασίες παραγωγής, όλες με χρήση του ίδιου μοντέλου εκτίμησης του θορύβου  $\hat{\boldsymbol{\epsilon}}_{\theta}$ . Από όλες τις διαφορετικές τιμές που μπορεί να λάβει η παράμετρος αυτή, δύο παρουσιάζουν ιδιαίτερο ενδιαφέρον. Πιο συγκεκριμένα:

- Όταν  $\sigma_t = \sqrt{(1 - \alpha_{t-1}) / (1 - \alpha_t)} \sqrt{1 - \alpha_t / \alpha_{t-1}}$  για κάθε  $t$ , η forward διαδικασία μετατρέπεται σε Μαρκοβιανή διαδικασία και ως εκ τούτου η διαδικασία παραγωγής αντιστοιχεί σε αυτή της περίπτωσης των DDPMs.
- Όταν  $\sigma_t = 0$  για κάθε  $t$ , η forward διαδικασία γίνεται πλήρως ντετερμινιστική, δοθέντων των  $\mathbf{x}_{t-1}$  και  $\mathbf{x}_t$ , με εξαίρεση την περίπτωση όπου  $t = 1$ . Το μοντέλο που προκύπτει σε αυτή την περίπτωση ονομάζεται denoising diffusion implicit model (DDIM), καθώς πρόκειται για ένα implicit πιθανοτικό μοντέλο [18], το οποίο όμως εκπαιδεύεται βάσει της ίδιας αντικειμενικής συνάρτησης με αυτή των DDPMs.

### 2.4.1 Επιτάχυνση Διαδικασίας Παραγωγής Δειγμάτων

Βάσει των όσων αναφέρθηκαν έως τώρα, η διαδικασία παραγωγής θεωρείται ως η προσέγγιση της αντίστροφης διαδικασίας. Αυτό σημαίνει ότι σε περίπτωση όπου η forward διαδικασία περιλαμβάνει  $T$  βήματα, για την παραγωγή ενός δείγματος απαιτούνται ομοίως  $T$  βήματα δειγματοληψίας. Παρόλα αυτά, σκοπός των DDIMs είναι η μείωση των βημάτων της διαδικασίας δειγματοληψίας.

Για το λόγο αυτό, θεωρείται η forward διαδικασία, η οποία δεν ορίζεται σε όλες τις επιμέρους παραμέτρους  $\mathbf{x}_{1:T}$ , αλλά σε ένα υποσύνολο  $\{\mathbf{x}_{\tau_1}, \dots, \mathbf{x}_{\tau_S}\}$  αυτών, όπου  $\tau$  είναι μία



αύξουσα υπακολουθία  $[1, \dots, T]$  μήκους  $S$ . Βάσει της υπακολουθίας αυτής, η κατανομή 2.87 παραγοντοποιείται ως εξής:

$$q_{\sigma, \tau}(\mathbf{x}_{1:T} | \mathbf{x}_0) = q_{\sigma, \tau}(\mathbf{x}_{\tau_S} | \mathbf{x}_0) \prod_{i=1}^S q_{\sigma, \tau}(\mathbf{x}_{\tau_{i-1}} | \mathbf{x}_{\tau_i}, \mathbf{x}_0) \prod_{i \in \bar{\tau}} q_{\sigma, \tau}(\mathbf{x}_i | \mathbf{x}_0), \quad (2.93)$$

όπου  $\bar{\tau} := \{1, \dots, T\} \setminus \tau$  το συμπληρωματικό του  $\tau$ . Ορίζονται τώρα τα ακόλουθα:

$$\begin{aligned} q_{\sigma, \tau}(\mathbf{x}_i | \mathbf{x}_0) &= \mathcal{N}(\sqrt{\alpha_i} \mathbf{x}_0, (1 - \alpha_i) \mathbf{I}) \quad \forall i \in \bar{\tau} \cup \{T\} \\ q_{\sigma, \tau}(\mathbf{x}_{\tau_{i-1}} | \mathbf{x}_{\tau_i}, \mathbf{x}_0) &= \mathcal{N}\left(\sqrt{\alpha_{\tau_{i-1}}} \mathbf{x}_0 + \sqrt{1 - \alpha_{\tau_{i-1}} - \sigma_{\tau_i}^2} \cdot \frac{\mathbf{x}_{\tau_i} - \sqrt{\alpha_{\tau_i}} \mathbf{x}_0}{\sqrt{1 - \alpha_{\tau_i}}}, \sigma_{\tau_i}^2 \mathbf{I}\right) \quad \forall i \in [S], \end{aligned} \quad (2.94)$$

όπου οι συντελεστές επιλέγονται ώστε να οδηγούν στις επιθυμητές περιθώριες κατανομές, σύμφωνα με τη σχέση,

$$q_{\sigma, \tau}(\mathbf{x}_{\tau_i} | \mathbf{x}_0) = \mathcal{N}(\sqrt{\alpha_{\tau_i}} \mathbf{x}_0, (1 - \alpha_{\tau_i}) \mathbf{I}) \quad \forall i \in [S] \quad (2.95)$$

Βάσει τώρα όλων των ανωτέρω, η διαδικασία παραγωγής ορίζεται ως εξής:

$$p_{\theta}(\mathbf{x}_{0:T}) := \underbrace{p_{\theta}(\mathbf{x}_T) \prod_{i=1}^S p_{\theta}(\mathbf{x}_{\tau_{i-1}} | \mathbf{x}_{\tau_i})}_{\text{παραγωγή δειγμάτων}} \times \underbrace{\prod_{i \in \bar{\tau}} p_{\theta}(\mathbf{x}_0 | \mathbf{x}_i)}_{\text{αντικειμενική συνάρτηση}}. \quad (2.96)$$

Στην ανωτέρω σχέση, μόνο ένα τμήμα χρησιμοποιείται για την παραγωγή δειγμάτων, ενώ οι δεσμευμένες κατανομές δίνονται από τις σχέσεις,

$$\begin{aligned} p_{\theta}(\mathbf{x}_{\tau_{i-1}} | \mathbf{x}_{\tau_i}) &= q_{\sigma, \tau}(\mathbf{x}_{\tau_{i-1}} | \mathbf{x}_{\tau_i}, f_{\theta}(\mathbf{x}_{\tau_{i-1}})) \quad \text{αν } i \in [S], i > 1 \\ p_{\theta}(\mathbf{x}_0 | \mathbf{x}_i) &= \mathcal{N}(f_{\theta}(\mathbf{x}_i), \sigma_i^2 \mathbf{I}) \quad \text{αλλιώς.} \end{aligned} \quad (2.97)$$

Η αντικειμενική συνάρτηση που χρησιμοποιείται τότε δίνεται από τη σχέση,

$$\begin{aligned} J &= \mathbb{E}_{\mathbf{x}_{0:T} \sim q_{\sigma, \tau}(\mathbf{x}_{0:T})} [\log q_{\sigma, \tau}(\mathbf{x}_{1:T} | \mathbf{x}_0) - \log p_{\theta}(\mathbf{x}_{0:T})] \\ &= \mathbb{E}_{\mathbf{x}_{0:T} \sim q_{\sigma, \tau}(\mathbf{x}_{0:T})} \left[ \sum_{i \in \bar{\tau}} D_{\text{KL}}(q_{\sigma, \tau}(\mathbf{x}_i | \mathbf{x}_0) \| p_{\theta}(\mathbf{x}_0 | \mathbf{x}_i)) \right. \\ &\quad \left. + \sum_{i=1}^L D_{\text{KL}}(q_{\sigma, \tau}(\mathbf{x}_{\tau_{i-1}} | \mathbf{x}_{\tau_i}, \mathbf{x}_0) \| p_{\theta}(\mathbf{x}_{\tau_{i-1}} | \mathbf{x}_{\tau_i})) \right], \end{aligned} \quad (2.98)$$

Τελικά, αποδεικνύεται ότι η αντικειμενική συνάρτηση της σχέσης 2.98 μπορεί να μετατραπεί σε μία ισοδύναμη αντικειμενική συνάρτηση της μορφής 2.43. Αυτό πρακτικά σημαίνει, ότι ένα μοντέλο μπορεί να εκπαιδευτεί βάσει ενός τυχαίου αριθμού βημάτων για τη forward διαδικασία και εν συνεχεία η λήψη νέων δειγμάτων να πραγματοποιηθεί χρησιμοποιώντας από ένα υποσύνολο των βημάτων αυτών. Κατά αυτό τον τρόπο επιταχύνεται σημαντικά η συνολική διαδικασία παραγωγής.

## Κεφάλαιο 3

### Text-to-Image Synthesis

---

Το πρόβλημα της σύνθεσης εικόνων (*image synthesis*) έχει προσελκύσει από τις απαρχές του το ενδιαφέρον αρκετών ερευνητών στον τομέα της υπολογιστικής όρασης. Ιδιαίτερα η σύνθεση ρεαλιστικών, υψηλής ανάλυσης εικόνων, οι οποίες αποτυπώνουν περίπλοκα φυσικά τοπία, αποτελεί μείζον αντικείμενο έρευνας, το οποίο ωστόσο είναι αρκετά απαιτητικό λόγω των υψηλών υπολογιστικών απαιτήσεών του. Για την αντιμετώπιση του προβλήματος αυτού έχουν προταθεί και αναπτυχθεί αρκετά μοντέλα, με χαρακτηριστικότερο παράδειγμα αυτό των GANs [19, 20, 21], τα οποία παρουσιάζουν πολλά υποσχόμενα αποτελέσματα, τα οποία ωστόσο περιορίζονται σε δεδομένα με σχετικά περιορισμένη μεταβλητότητα, γεγονός το οποίο οφείλεται στην ανταγωνιστική φύση της διαδικασίας που ακολουθείται για την εκπαίδευσή τους. Τα τελευταία χρόνια, η ανάπτυξη και η χρήση των μοντέλων διάχυσης έχει παρουσιάσει θεαματικά αποτελέσματα στη σύνθεση ρεαλιστικών εικόνων, υπερβαίνοντας τα προβλήματα που παρουσιάζονται στην περίπτωση των GANs. Πιο συγκεκριμένα, όντας likelihood-based μοντέλα, δεν παρουσιάζουν το φαινόμενο του mode-collapse και τις αστάθειες κατά την εκπαίδευση που παρατηρούνται στην περίπτωση των GANs, ενώ μέσω του διαμοιρασμού των παραμέτρων, είναι σε θέση να μοντελοποιούν πολύ σύνθετες κατανομές, χωρίς να απαιτείται η χρήση δισεκατομμυρίων παραμέτρων.

Μία επέκταση του προβλήματος της σύνθεσης εικόνων, είναι η καθοδηγούμενη από κειμενική περιγραφή σύνθεση εικόνων (*text-to-image synthesis*), δηλαδή η σύνθεση ρεαλιστικών εικόνων, των οποίων οι απεικονίσεις προκύπτουν βάσει περιγραφών, οι οποίες παρατίθενται υπό τη μορφή κειμένου. Το πρόβλημα αυτό είναι ακόμη πιο απαιτητικό από την απλή περίπτωση της σύνθεσης εικόνων, καθώς πλέον, εισάγονται στη διαδικασία της σύνθεσης επιπλέον συνθήκες. Για την αντιμετώπιση του εν λόγω προβλήματος, όπως και στην απλή περίπτωση, έχουν αναπτυχθεί τόσο μοντέλα GANs όσο και μοντέλα διάχυσης, με τα πρώτα εξ αυτών να παρουσιάζουν τα προβλήματα που προαναφέρθηκαν. Στα πλαίσια της παρούσας εργασίας, θα εξεταστεί η περίπτωση των μοντέλων διάχυσης και συγκεκριμένα θα αναλυθεί η λειτουργία ενός εκ των πιο γνωστών μοντέλων για την καθοδηγούμενη από κειμενική περιγραφή σύνθεση εικόνων, του **Stable Diffusion** [22].

### 3.1 Καθοδήγηση Μοντέλων Διάχυσης

Έως το σημείο αυτό, το θεωρητικό υπόβαθρο το οποίο έχει θεμελιωθεί, αφορά στη λειτουργία μοντέλων διάχυσης, τα οποία συνθέτουν εικόνες χωρίς τον περιορισμό και την καθοδήγηση εξωτερικών συνθηκών (*unconditional diffusion models*). Παρά ταύτα, τα μοντέλα διάχυσης, όπως και άλλοι τύποι παραγωγικών μοντέλων [23] είναι ικανά στη γενική περίπτωση, να μοντελοποιούν δεσμευμένες κατανομές.

Έστω λοιπόν ότι ένα δοθέν δείγμα  $\mathbf{x}_0$ , προέρχεται από μία δεσμευμένη κατανομή  $p(\mathbf{x}|\mathbf{y})$ , όπου  $\mathbf{y}$  μία συνθήκη, η οποία μπορεί να αναφέρεται σε κείμενο [24], σε σημασιολογικούς χάρτες [25] κτλ. Σύμφωνα με τις βασικές αρχές λειτουργίας των DDPMs, προστίθεται διαδοχικά Γκαουσιανός θόρυβος στο αρχικό αυτό δείγμα, βάσει της σχέσης,

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}) = \mathcal{N}\left(\mathbf{x}_t|\mathbf{y}; \boldsymbol{\mu}_t = \sqrt{1 - \beta_t}\mathbf{x}_{t-1}|\mathbf{y}, \boldsymbol{\Sigma}_t = \beta_t\mathbf{I}\right), \text{ για } t = \{1, \dots, T\}. \quad (3.1)$$

Στην περίπτωση αυτή, το ενθόρυβο δείγμα  $\mathbf{x}_t$  μπορεί να εκφραστεί συναρτήσει του αρχικού δείγματος  $\mathbf{x}_0$  ως εξής:

$$q(\mathbf{x}_t|\mathbf{x}_0, \mathbf{y}) := \mathcal{N}\left(\mathbf{x}_t|\mathbf{y}; \sqrt{\bar{\alpha}_t}\mathbf{x}_0|\mathbf{y}, (1 - \bar{\alpha}_t)\mathbf{I}\right), \quad (3.2)$$

όπου ως γνωστόν  $\alpha_t = 1 - \beta_t$  και  $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$ . Σκοπός επομένως είναι η προσέγγιση της αντίστροφης διαδικασίας διάχυσης, η οποία και περιγράφεται από την κάτωθι σχέση:

$$p_\theta(\mathbf{x}_{0:T}|\mathbf{y}) = \prod_{t=1}^T p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{y}). \quad (3.3)$$

Από την ανωτέρω σχέση εύκολα γίνεται αντιληπτό ότι η συνθήκη υπεισέρχεται σε κάθε βήμα της αντίστροφης διαδικασίας διάχυσης. Αρκεί επομένως να βρεθεί ένας τρόπος, ώστε σε κάθε επιμέρους βήμα  $t$  της διαδικασίας αυτής να γίνεται δειγματοληψία από την κατανομή  $p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{y})$ .

Στις επόμενες ενότητες γίνεται αναφορά σε δύο οικογένειες μεθόδων, οι οποίες αποσκοπούν στην αντιμετώπιση του προβλήματος της καθοδηγούμενης σύνθεσης: η πρώτη εξ αυτών είναι η μέθοδος *classifier guidance* και η δεύτερη η μέθοδος *classifier free guidance*. Στα πλαίσια της επακόλουθης ανάλυσης και για λόγους ευκολίας και συμμόρφωσης με τις αρχικές υλοποιήσεις, η συνθήκη  $\mathbf{y}$  θεωρείται ότι αναφέρεται στις επιθυμητές ετικέτες των κλάσεων (*class label*) των παραγόμενων εικόνων, χωρίς ωστόσο αυτό να σημαίνει ότι δεν μπορεί να εκφράζει και κειμενικές περιγραφές.

#### 3.1.1 Classifier Guidance

Οι *Sohl et al.* [26] αρχικά και έπειτα οι *Dhariwal et al.* [27], έδειξαν ότι μπορεί να χρησιμοποιηθεί ένα δεύτερο μοντέλο ταξινομητή  $p_\phi(\mathbf{y}|\mathbf{x}_t, t)$  προκειμένου να καθοδηγηθεί η διαδικασία της διάχυσης προς την επιθυμητή κλάση  $\mathbf{y}$ .

Πιο αναλυτικά, έστω ένα μοντέλο διάχυσης, με αντίστροφη διαδικασία διάχυσης η οποία δεν υπόκειται σε κάποια εξωτερική συνθήκη  $p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})$ . Προκειμένου η διαδικασία αυτή να λάβει υπόψιν τη συνθήκη  $\mathbf{y}$ , αρκεί όπως αναφέρθηκε και ανωτέρω, να γίνει δειγματοληψία σε κάθε επιμέρους βήμα αποθορυβοποίησης, από την κατανομή,

$$p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{y}) = p_{\theta, \phi}(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{y}) = Z p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1}) p_\phi(\mathbf{y}|\mathbf{x}_t), \quad (3.4)$$

όπου  $Z$  μία σταθερά κανονικοποίησης. Σημαντική παρατήρηση είναι ότι η Σχ.3.4 δεν προκύπτει άμεσα, με αποτέλεσμα να μην είναι προφανής. Ως εκ τούτου και για λόγους πληρότητας, θα παρουσιαστεί εν συντομία η απόδειξή της.

Αρχικά, ορίζεται μία υπό συνθήκη Μαρκοβιανή διαδικασία θορύβου  $\hat{q}$ , η οποία είναι παρόμοια με την αντίστοιχη διαδικασία  $q$  και γίνεται η υπόθεση, ότι η κατανομή των ετικετών των κλάσεων  $\hat{q}(\mathbf{y}|\mathbf{x}_0)$  είναι εκ των προτέρων γνωστή και διαθέσιμη για κάθε δείγμα. Επομένως, ισχύουν τα ακόλουθα:

$$\begin{aligned} \hat{q}(\mathbf{x}_0) &:= q(\mathbf{x}_0) \\ \hat{q}(\mathbf{y}|\mathbf{x}_0) &:= \text{Γνωστές ετικέτες για κάθε δείγμα} \\ \hat{q}(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{y}) &:= q(\mathbf{x}_{t+1}|\mathbf{x}_t) \\ \hat{q}(\mathbf{x}_{1:T}|\mathbf{x}_0, \mathbf{y}) &:= \prod_{t=1}^T \hat{q}(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}) \end{aligned}$$

Παρόλου που η διαδικασία θορύβου  $\hat{q}$  έχει οριστεί βάσει της συνθήκης  $\mathbf{y}$ , μπορεί να αποδειχθεί ότι συμπεριφέρεται ακριβώς σαν τη διαδικασία  $q$ , όταν αφαιρείται η επίδραση της συνθήκης  $\mathbf{y}$ . Στο πλαίσιο της παρατήρησης αυτής, ισχύει,

$$\begin{aligned} \hat{q}(\mathbf{x}_{t+1}|\mathbf{x}_t) &= \int_{\mathbf{y}} \hat{q}(\mathbf{x}_{t+1}, \mathbf{y}|\mathbf{x}_t) d\mathbf{y} \\ &= \int_{\mathbf{y}} \hat{q}(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{y}) \hat{q}(\mathbf{y}|\mathbf{x}_t) d\mathbf{y} \\ &= \int_{\mathbf{y}} q(\mathbf{x}_{t+1}|\mathbf{x}_t) \hat{q}(\mathbf{y}|\mathbf{x}_t) d\mathbf{y} \\ &= q(\mathbf{x}_{t+1}|\mathbf{x}_t) \int_{\mathbf{y}} \hat{q}(\mathbf{y}|\mathbf{x}_t) d\mathbf{y} \\ &= q(\mathbf{x}_{t+1}|\mathbf{x}_t) \\ &= \hat{q}(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{y}). \end{aligned} \quad (3.5)$$

Βάσει τώρα παρόμοιας λογικής, η από κοινού κατανομή  $\hat{q}(\mathbf{x}_{1:T}|\mathbf{x}_0)$  θα ισούται με:

$$\begin{aligned} \hat{q}(\mathbf{x}_{1:T}|\mathbf{x}_0) &= \int_{\mathbf{y}} \hat{q}(\mathbf{x}_{1:T}, \mathbf{y}|\mathbf{x}_0) d\mathbf{y} \\ &= \int_{\mathbf{y}} \hat{q}(\mathbf{y}|\mathbf{x}_0) \hat{q}(\mathbf{x}_{1:T}|\mathbf{x}_0, \mathbf{y}) d\mathbf{y} \end{aligned}$$

$$\begin{aligned}
&= \int_{\mathbf{y}} \hat{q}(\mathbf{y}|\mathbf{x}_0) \prod_{t=1}^T \hat{q}(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}) d\mathbf{y} \\
&= \int_{\mathbf{y}} \hat{q}(\mathbf{y}|\mathbf{x}_0) \prod_{t=1}^T q(\mathbf{x}_t|\mathbf{x}_{t-1}) d\mathbf{y} \\
&= \prod_{t=1}^T q(\mathbf{x}_t|\mathbf{x}_{t-1}) \int_{\mathbf{y}} \hat{q}(\mathbf{y}|\mathbf{x}_0) d\mathbf{y} \\
&= \prod_{t=1}^T q(\mathbf{x}_t|\mathbf{x}_{t-1}) \\
&= q(\mathbf{x}_{1:T}|\mathbf{x}_0). \tag{3.6}
\end{aligned}$$

Χρησιμοποιώντας τώρα την εξίσωση 3.6, η κατανομή  $\hat{q}(\mathbf{x}_t)$  προκύπτει ίση με,

$$\begin{aligned}
\hat{q}(\mathbf{x}_t) &= \int_{\mathbf{x}_{0:t-1}} \hat{q}(\mathbf{x}_0, \dots, \mathbf{x}_t) d\mathbf{x}_{0:t-1} \\
&= \int_{\mathbf{x}_{0:t-1}} \hat{q}(\mathbf{x}_0) \hat{q}(\mathbf{x}_1, \dots, \mathbf{x}_t|\mathbf{x}_0) d\mathbf{x}_{0:t-1} \\
&= \int_{\mathbf{x}_{0:t-1}} q(\mathbf{x}_0) q(\mathbf{x}_1, \dots, \mathbf{x}_t|\mathbf{x}_0) d\mathbf{x}_{0:t-1} \\
&= q(\mathbf{x}_t). \tag{3.7}
\end{aligned}$$

Αξιοποιώντας τις ιδιότητες  $\hat{q}(\mathbf{x}_t) = q(\mathbf{x}_t)$  και  $\hat{q}(\mathbf{x}_{t+1}|\mathbf{x}_t) = q(\mathbf{x}_{t+1}|\mathbf{x}_t)$  και σύμφωνα με τον κανόνα του Bayes, είναι πλέον τετριμμένο να αποδειχθεί ότι για την μη δεσμευμένη αντίστροφη διαδικασία ισχύει  $\hat{q}(\mathbf{x}|\mathbf{x}_{t+1}) = q(\mathbf{x}|\mathbf{x}_{t+1})$ .

Μία σημαντική παρατήρηση σχετικά με την κατανομή  $\hat{q}$ , η οποία και θα αξιοποιηθεί στη συνέχεια, είναι ότι, αν θεωρηθεί μία θορυβώδης κατανομή ταξινόμησης  $\hat{q}(\mathbf{y}|\mathbf{x}_t)$ , μπορεί να αποδειχθεί ότι η τελευταία αυτή κατανομή είναι ανεξάρτητη του δείγματος  $\mathbf{x}_{t+1}$ . Δηλαδή,

$$\begin{aligned}
\hat{q}(\mathbf{y}|\mathbf{x}_t, \mathbf{x}_{t+1}) &= \hat{q}(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{y}) \frac{\hat{q}(\mathbf{y}|\mathbf{x}_t)}{\hat{q}(\mathbf{x}_{t+1}|\mathbf{x}_t)} \\
&= \hat{q}(\mathbf{x}_{t+1}|\mathbf{x}_t) \frac{\hat{q}(\mathbf{y}|\mathbf{x}_t)}{\hat{q}(\mathbf{x}_{t+1}|\mathbf{x}_t)} \\
&= \hat{q}(\mathbf{y}|\mathbf{x}_t). \tag{3.8}
\end{aligned}$$

Συνδυάζοντας τώρα όλα τα παραπάνω, η υπό συνθήκη αντίστροφη διαδικασία ορίζεται ως εξής:

$$\begin{aligned}
\hat{q}(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{y}) &= \frac{\hat{q}(\mathbf{x}_t, \mathbf{x}_{t+1}, \mathbf{y})}{\hat{q}(\mathbf{x}_{t+1}, \mathbf{y})} \\
&= \frac{\hat{q}(\mathbf{x}_t, \mathbf{x}_{t+1}, \mathbf{y})}{\hat{q}(\mathbf{y}|\mathbf{x}_{t+1}) \hat{q}(\mathbf{x}_{t+1})}
\end{aligned}$$

$$\begin{aligned}
&= \frac{\hat{q}(\mathbf{x}_t|\mathbf{x}_{t+1})\hat{q}(\mathbf{y}|\mathbf{x}_t, \mathbf{x}_{t+1})\hat{q}(\mathbf{x}_{t+1})}{\hat{q}(\mathbf{y}|\mathbf{x}_{t+1})\hat{q}(\mathbf{x}_{t+1})} \\
&= \frac{\hat{q}(\mathbf{x}_t|\mathbf{x}_{t+1})\hat{q}(\mathbf{y}|\mathbf{x}_t, \mathbf{x}_{t+1})}{\hat{q}(\mathbf{y}|\mathbf{x}_{t+1})} \\
&= \frac{\hat{q}(\mathbf{x}_t|\mathbf{x}_{t+1})\hat{q}(\mathbf{y}|\mathbf{x}_t)}{\hat{q}(\mathbf{y}|\mathbf{x}_{t+1})} \\
&= \frac{q(\mathbf{x}_t|\mathbf{x}_{t+1})\hat{q}(\mathbf{y}|\mathbf{x}_t)}{\hat{q}(\mathbf{y}|\mathbf{x}_{t+1})} \tag{3.9}
\end{aligned}$$

Ο όρος  $\hat{q}(\mathbf{y}|\mathbf{x}_{t+1})$  της εξίσωσης 3.9 μπορεί να αντιμετωπιστεί ως σταθερός, εφόσον δεν εξαρτάται από το τυχόν δείγμα  $\mathbf{x}_t$ . Προκειμένου επομένως να εισαχθεί η καθοδήγηση στη διαδικασία της διάχυσης, θα πρέπει να γίνει δειγματοληψία από την κατανομή

$$Zq(\mathbf{x}_t|\mathbf{x}_{t+1})\hat{q}(\mathbf{y}|\mathbf{x}_t), \tag{3.10}$$

όπου  $Z$  η σταθερά κανονικοποίησης. Εφόσον υπάρχει ήδη διαθέσιμο ένα unconditional μοντέλο διάχυσης, η προσέγγιση  $p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})$  της κατανομής  $q(\mathbf{x}_t|\mathbf{x}_{t+1})$  είναι γνωστή. Επομένως το μόνο που μένει, είναι η εκπαίδευση ενός ταξινομητή  $p_\phi(\mathbf{y}|\mathbf{x}_t)$ , χρησιμοποιώντας ενθόρυβες εικόνες  $\mathbf{x}_t$ , οι οποίες και λαμβάνονται μέσω δειγματοληψίας από την κατανομή  $q(\mathbf{x}_t)$ .

Η κατανομή της σχέσης 3.10 είναι ιδιαίτερος δύσκολο, αν όχι ακατόρθωτο να προσδιοριστεί επακριβώς, ώστε εν συνεχεία κάποιος να προβεί σε δειγματοληψία από αυτή. Παρά ταύτα, οι *Sohl-Dickstein et al.* [26] έδειξαν ότι η κατανομή αυτή μπορεί να προσεγγιστεί ως μία ελαφρώς διαταραγμένη Γκαουσιανή κατανομή. Ας σημειωθεί στο σημείο αυτό, ότι ένα μοντέλο διάχυσης προβλέπει το δείγμα  $\mathbf{x}_t$ , βάσει του δείγματος  $\mathbf{x}_{t+1}$  του επόμενου χρονικού βήματος, δειγματοληπώντας από μία Γκαουσιανή κατανομή με,

$$p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1}) = \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta) \tag{3.11}$$

$$\log p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1}) = -\frac{1}{2}(\mathbf{x}_t - \boldsymbol{\mu}_\theta)^T \boldsymbol{\Sigma}_\theta^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_\theta) + C, \tag{3.12}$$

όπου στις παραπάνω σχέσεις εννοείται ότι  $\boldsymbol{\mu}_\theta = \boldsymbol{\mu}_\theta(\mathbf{x}_t, t)$  και  $\boldsymbol{\Sigma}_\theta = \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)$ .

Αν τώρα υποθεθεί ότι ο όρος  $\log p_\theta(\mathbf{y}|\mathbf{x}_t)$  έχει χαμηλή κυρτότητα σε σχέση με τον όρο της συνδιακύμανσης  $\boldsymbol{\Sigma}_\theta^{-1}$ , μπορεί να προσεγγιστεί με χρήση του αναπτύγματος Taylor γύρω από το  $\mathbf{x}_t = \boldsymbol{\mu}_\theta$  σύμφωνα με τη σχέση,

$$\begin{aligned}
\log p_\phi(\mathbf{y}|\mathbf{x}_t) &\approx \log p_\phi(\mathbf{y}|\mathbf{x}_t)|_{\mathbf{x}_t=\boldsymbol{\mu}_\theta} + (\mathbf{x}_t - \boldsymbol{\mu}_\theta)^T \nabla_{\mathbf{x}_t} \log p_\phi(\mathbf{y}|\mathbf{x}_t)|_{\mathbf{x}_t=\boldsymbol{\mu}_\theta} \\
&= (\mathbf{x}_t - \boldsymbol{\mu}_\theta)^T \mathbf{g} + C_1, \tag{3.13}
\end{aligned}$$

όπου  $\mathbf{g} = \nabla_{\mathbf{x}_t} \log p_\phi(\mathbf{y}|\mathbf{x}_t)|_{\mathbf{x}_t=\boldsymbol{\mu}_\theta}$  και  $C_1$  σταθερός όρος.

Χρησιμοποιώντας το ανάπτυγμα αυτό προκύπτει ότι,

$$\begin{aligned}
\log(p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})p_\phi(\mathbf{y}|\mathbf{x}_t)) &\approx -\frac{1}{2}(\mathbf{x}_t - \boldsymbol{\mu}_\theta)^T \boldsymbol{\Sigma}_\theta^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_\theta) + (\mathbf{x}_t - \boldsymbol{\mu}_\theta) \mathbf{g} + C_2 \\
&= -\frac{1}{2}(\mathbf{x}_t - \boldsymbol{\mu}_\theta - \boldsymbol{\Sigma}_\theta \mathbf{g})^T \boldsymbol{\Sigma}_\theta^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_\theta - \boldsymbol{\Sigma}_\theta \mathbf{g}) + \frac{1}{2} \mathbf{g}^T \boldsymbol{\Sigma}_\theta \mathbf{g} + C_2 \\
&= -\frac{1}{2}(\mathbf{x}_t - \boldsymbol{\mu}_\theta - \boldsymbol{\Sigma}_\theta \mathbf{g})^T \boldsymbol{\Sigma}_\theta^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_\theta - \boldsymbol{\Sigma}_\theta \mathbf{g}) + C_3 \\
&= \log p(\mathbf{z}) + C_4, \quad \mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}_\theta + \boldsymbol{\Sigma}_\theta \mathbf{g}, \boldsymbol{\Sigma}_\theta). \tag{3.14}
\end{aligned}$$

Στην τελευταία σχέση, όρος  $C_4$ , όντας σταθερός, αντιστοιχεί στη σταθερά κανονικοποίησης  $Z$  και ως εκ τούτου μπορεί να παραληφθεί με ασφάλεια. Έτσι, συμπεραίνεται ότι ο υπό συνθήκη τελεστής μετάβασης της σχέσης 3.4 μπορεί να προσεγγιστεί μέσω ενός Γκαουσιανού πυρήνα, παρόμοιου με αυτού του αντίστοιχου unconditional τελεστή, με τη μόνη διαφορά, ότι η μέση τιμή του είναι μετατοπισμένη κατά παράγοντα  $\boldsymbol{\Sigma}_\theta \mathbf{g}$ . Έπειτα από αρκετούς πειραματισμούς οι *Dhariwal et al.* [27] διαπίστωσαν ότι είναι απαραίτητη η προσθήκη ενός επιπλέον όρου κλιμάκωσης  $s$ , ο οποίος κλιμακώνει την επίδραση των παραγώγων του ταξινομητή. Επομένως, η τελική δειγματοληψία πραγματοποιείται από μία Γκαουσιανή κατανομή της μορφής  $\mathcal{N}(\boldsymbol{\mu}_\theta + s\boldsymbol{\Sigma}_\theta \mathbf{g}, \boldsymbol{\Sigma}_\theta)$ . Όσο μεγαλύτερη του 1 είναι η τιμή του παράγοντα  $s$ , τόσο μεγαλύτερη είναι και η επίδραση της εξωτερικής συνθήκης  $\mathbf{y}$  στη διαδικασία της αντίστροφης διάχυσης. Στον Αλγόριθμο 3.1 παρουσιάζεται ο αλγόριθμος δειγματοληψίας για τη διαδικασία της αντίστροφης διάχυσης, όταν αυτή καθοδηγείται βάσει κάποιας συνθήκης  $\mathbf{y}$ .

---

**Αλγόριθμος 3.1:** Classifier guided diffusion sampling.

---

**Require:** class label  $\mathbf{y}$ , gradient scaling factor  $s$

- 1:  $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  - 2: **for**  $t = T, \dots, 1$  **do**
  - 3:      $\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta \leftarrow \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)$
  - 4:      $\mathbf{x}_{t-1} \sim \mathcal{N}(\boldsymbol{\mu}_\theta + \boldsymbol{\Sigma}_\theta \mathbf{g}, \boldsymbol{\Sigma}_\theta)$
  - 5: **end for**
  - 6: **return**  $\mathbf{x}_0$
- 

### Καθοδηγούμενη Δειγματοληψία για DDIM

Η διαδικασία της καθοδηγούμενης δειγματοληψίας που περιγράφηκε έως τώρα, είναι έγκυρη μόνο σε περιπτώσεις όπου χρησιμοποιούνται στοχαστικές διαδικασίες για την υλοποίηση της αντίστροφης διάχυσης και δεν μπορεί να εφαρμοστεί άμεσα σε ντετερμινιστικές μεθόδους δειγματοληψίας, όπως τα DDIMs. Χρειάζεται επομένως κατάλληλη τροποποίηση.

Προς την κατεύθυνση αυτή, έστω ένα μοντέλο  $\hat{\epsilon}_\theta(\mathbf{x}_t, t)$  το οποίο έχει εκπαιδευτεί ώστε να παρέχει μία εκτίμηση του διανύσματος θορύβου που έχει προστεθεί στο δείγμα  $\mathbf{x}_t$  σε

κάποιο επίπεδο θορύβου  $t$ . Η εκτίμηση αυτή συνδέεται με την αντίστοιχη score function μέσω της σχέσης, [8]

$$\nabla_{\mathbf{x}_t} \log p_{\theta}(\mathbf{x}_t) = -\frac{1}{\sqrt{1 - \bar{\alpha}_t}} \hat{\boldsymbol{\epsilon}}_{\theta}(\mathbf{x}_t, t) \quad (3.15)$$

Αντικαθιστώντας τη σχέση 3.15 στη score function της κατανομής  $p_{\theta}(\mathbf{x}_t)p_{\phi}(\mathbf{y}|\mathbf{x}_t)$  προκύπτει το εξής:

$$\begin{aligned} \nabla_{\mathbf{x}_t} \log (p_{\theta}(\mathbf{x}_t)p_{\phi}(\mathbf{y}|\mathbf{x}_t)) &= \nabla_{\mathbf{x}_t} \log p_{\theta}(\mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log p_{\phi}(\mathbf{y}|\mathbf{x}_t) \\ &= -\frac{1}{\sqrt{1 - \bar{\alpha}_t}} \hat{\boldsymbol{\epsilon}}_{\theta}(\mathbf{x}_t, t)(\mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log p_{\phi}(\mathbf{y}|\mathbf{x}_t) \end{aligned} \quad (3.16)$$

Τελικά, ορίζεται ένα ανανεωμένο διάνυσμα εκτίμησης θορύβου  $\tilde{\boldsymbol{\epsilon}}_{\theta}(\mathbf{x}_t, t)$ , το οποίο εκφράζει το score της από κοινού κατανομής:

$$\tilde{\boldsymbol{\epsilon}}_{\theta}(\mathbf{x}_t, t) = \hat{\boldsymbol{\epsilon}}_{\theta}(\mathbf{x}_t, t) - \sqrt{1 - \bar{\alpha}_t} \nabla_{\mathbf{x}_t} p_{\phi}(\mathbf{y}|\mathbf{x}_t) \quad (3.17)$$

Η ανανεωμένη αυτή εκτίμηση μπορεί να χρησιμοποιηθεί για τη ντετερμινιστική δειγματοληψία με DDIM, όπως περιγράφεται και στον Αλγόριθμο 3.2.

---

**Αλγόριθμος 3.2:** Classifier guided DDIM sampling.

---

**Require:** class label  $\mathbf{y}$ , gradient scaling factor  $s$

- 1:  $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, I)$
  - 2: **for**  $t = T, \dots, 1$  **do**
  - 3:      $\tilde{\boldsymbol{\epsilon}}_{\theta}(\mathbf{x}_t, t) \leftarrow \hat{\boldsymbol{\epsilon}}_{\theta}(\mathbf{x}_t, t) - \sqrt{1 - \bar{\alpha}_t} \nabla_{\mathbf{x}_t} p_{\phi}(\mathbf{y}|\mathbf{x}_t)$
  - 4:      $\mathbf{x}_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}} \left( \frac{\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \tilde{\boldsymbol{\epsilon}}_{\theta}(\mathbf{x}_t, t)}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1 - \bar{\alpha}_{t-1}} \tilde{\boldsymbol{\epsilon}}_{\theta}(\mathbf{x}_t, t)$
  - 5: **end for**
  - 6: **return**  $\mathbf{x}_0$
- 

### 3.1.2 Classifier-Free Guidance

Η μέθοδος classifier guidance που περιγράφηκε στην προηγούμενη ενότητα επιτυγχάνει να οδηγήσει τη διαδικασία της αντίστροφης διάχυσης, βάσει της εξωτερικής συνθήκης  $\mathbf{y}$ , βελτιώνοντας αισθητά τα τελικά αποτελέσματα. Παρά την αποτελεσματικότητά της ωστόσο, η μέθοδος αυτή παρουσιάζει ένα σημαντικότατο μειονέκτημα, το οποίο δεν είναι άλλο από την άμεση εξάρτησή της από την ύπαρξη ενός επιπλέον μοντέλου ταξινόμησης. Η ύπαρξη του μοντέλου αυτού περιπλέκει σημαντικά τη διαδικασία της εκπαίδευσης των μοντέλων διάχυσης, καθώς απαιτεί την εκπαίδευση ενός εξωτερικού ταξινομητή, χρησιμοποιώντας θορυβώδη δείγματα, με αποτέλεσμα να μην καθίσταται δυνατή η χρήση ενός προεκπαιδευμένου μοντέλου. Είναι επομένως σκόπιμο να μελετηθεί το κατά πόσο η διαδικασία της καθοδήγησης μπορεί να υλοποιηθεί χωρίς την ύπαρξη κάποιου εξωτερικού ταξινομητή.



Βασιζόμενοι στην ανάγκη αυτή, οι *Ho et al.* [28] προτείνουν μία νέα μέθοδο για την καθοδήγηση της διαδικασίας της αντίστροφης διάχυσης, η οποία αποδεδειγμένη από τη χρήση των παραγώγων του μοντέλου ταξινόμησης. Η μέθοδος αυτή καλείται *classifier-free guidance* και συνίσταται στην τροποποίηση της εκτίμησης του θορύβου κατά τέτοιο τρόπο, ώστε να έχει την ίδια επίδραση με αυτή της περίπτωσης του classifier guidance, χωρίς ωστόσο να απαιτείται η ύπαρξη κάποιου ταξινομητή.

Πιο συγκεκριμένα, οι *Ho et al.* αντί να εκπαιδεύσουν ένα ξεχωριστό μοντέλο ταξινόμησης, επιλέγουν να εκπαιδεύσουν ένα unconditional μοντέλο διάχυσης  $p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1})$ , το οποίο παραμετροποιείται μέσω ενός μοντέλου εκτίμησης score  $\hat{\epsilon}_\theta(\mathbf{x}_t, t)$ . Συγχρόνως, εκπαιδεύουν και ένα conditional μοντέλο διάχυσης  $p_\theta(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{y})$ , το οποίο κατά αντίστοιχο τρόπο παραμετροποιείται μέσω ενός μοντέλου εκτίμησης score  $\hat{\epsilon}_\theta(\mathbf{x}_t, \mathbf{y}, t)$ . Τα δύο μοντέλα υλοποιούνται από κοινού από ένα και μόνο νευρωνικό δίκτυο, το οποίο στην περίπτωση του unconditional μοντέλου δέχεται ως συνθήκη (κλάση) την κενή ακολουθία  $\emptyset$ , δηλαδή  $\hat{\epsilon}_\theta(\mathbf{x}_t) = \hat{\epsilon}_\theta(\mathbf{x}_t, \mathbf{y} = \emptyset, t)$ . Η εκπαίδευση των δύο μοντέλων πραγματοποιείται επίσης από κοινού, με την κενή ακολουθία να επιλέγεται ως συνθήκη εισόδου με πιθανότητα  $p_{uncond}$ , η οποία και λειτουργεί εν είδει υπερπαραμέτρου. Στον Αλγόριθμο 3.3 παρουσιάζεται η διαδικασία της από κοινού εκπαίδευσης των δύο μοντέλων.

---

**Αλγόριθμος 3.3:** Joint training a diffusion model with classifier-free guidance.

---

**Require:**  $p_{uncond}$  : probability of unconditional training

- 1: **repeat**
  - 2:      $(\mathbf{x}_0, \mathbf{y}) \sim p(\mathbf{x}_0|\mathbf{y})$  ▷ Sample data with condition
  - 3:      $\mathbf{y} \leftarrow \emptyset$  with probability  $p_{uncond}$
  - 4:      $t \sim p(t)$  ▷ Sample log SNR value
  - 5:      $\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \epsilon, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  - 6:     Take gradient step on  $\nabla_\theta \|\hat{\epsilon}_\theta(\mathbf{x}_t, \mathbf{y}, t) - \epsilon\|^2$
  - 7: **until** converged
- 

Μετά την ολοκλήρωση της από κοινού εκπαίδευσης των δύο μοντέλων, η δειγματοληψία πραγματοποιείται χρησιμοποιώντας τον κάτωθι γραμμικό συνδυασμό των conditional και unconditional εκτιμήσεων των scores:

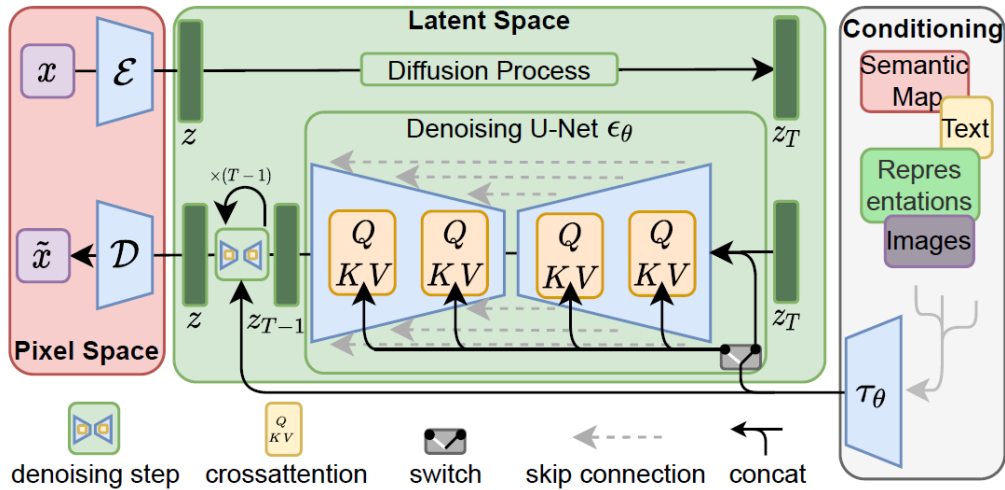
$$\begin{aligned} \tilde{\epsilon}_\theta(\mathbf{x}_t, \mathbf{y}, t) &= (1 + s)\hat{\epsilon}_\theta(\mathbf{x}_t, \mathbf{y}, t) - s\hat{\epsilon}_\theta(\mathbf{x}_t, t) \\ &= \hat{\epsilon}_\theta(\mathbf{x}_t, t) + s \cdot (\hat{\epsilon}_\theta(\mathbf{x}_t, \mathbf{y}, t) - \hat{\epsilon}_\theta(\mathbf{x}_t, t)). \end{aligned} \quad (3.18)$$

Από την τελευταία σχέση καθίσταται προφανές, ότι η ανανεωμένη εκδοχή της εκτίμησης του θορύβου δεν περιλαμβάνει υπολογισμό παραγώγων κάποιου ταξινομητή. Προφανώς, για  $s = 0$  χρησιμοποιείται μόνο το unconditional τμήμα του κοινού μοντέλου, ενώ καθώς η τιμή του  $s$  αυξάνεται, αυξάνεται και η επίδραση της συνθήκης. Συνολικά, η μέθοδος καθοδήγησης αυτή παρουσιάζει δύο σημαντικά πλεονεκτήματα:

- Χρησιμοποιεί μόνο ένα μοντέλο για να καθοδηγήσει τη διαδικασία της διάχυσης.
- Απλοποιεί σημαντικά την καθοδήγηση όταν η συνθήκη εισόδου εκφράζει πληροφορίες, οι οποίες είναι δύσκολο να προβλεφθούν μέσω ενός ταξινομητή (όπως για παράδειγμα οι κειμενικές περιγραφές).

## 3.2 Stable Diffusion

Το Stable Diffusion πρόκειται για ένα *latent* μοντέλο διάχυσης, το οποίο αναπτύχθηκε από τους *Rombach et al.* [22] με σκοπό την παραγωγή εικόνων υψηλής ευκρίνειας βάσει αντίστοιχων κειμενικών περιγραφών. Η συνολική αρχιτεκτονική του εν λόγω μοντέλου παρουσιάζεται στο Σχ.3.1.



Σχήμα 3.1: Αρχιτεκτονική Stable Diffusion [22].

Βάσει του σχήματος αυτού, η αρχιτεκτονική του Stable Diffusion περιλαμβάνει τα ακόλουθα τρία βασικά τμήματα:

1. **Autoencoder**, για τη μείωση της διαστατικότητας των δεδομένων και τη δημιουργία λανθανουσών (latent) αναπαραστάσεων.
2. **Denoising U-Net**, για την αφαίρεση του τυχαίου θορύβου που προστίθεται στα δεδομένα κατά τη διαδικασία της διάχυσης.
3. **Text Encoder**, για την επεξεργασία και τη μετατροπή της δοθείσης κειμενικής περιγραφής σε κατάλληλη embedding αναπαράσταση.

Στις ακόλουθες ενότητες θα γίνει ξεχωριστή αναφορά στα επιμέρους αυτά τμήματα της αρχιτεκτονικής του *Stable Diffusion*, προκειμένου να γίνει κατανοητός ο τρόπος λειτουργίας του.

### 3.2.1 Autoencoder (VAE)

Το πρώτο εκ των βασικών τμημάτων της αρχιτεκτονικής του Stable Diffusion είναι αυτό του *Autoencoder*. Καθώς η εφαρμογή των διαδικασιών διάχυσης και αποθορυβοποίησης των μοντέλων διάχυσης σε επίπεδο pixel είναι υπερβολικά κοστοβόρα και χρονοβόρα, το μοντέλο Stable Diffusion κάνει χρήση ενός *Variational Autoencoder*, προκειμένου να δημιουργηθούν αντιπροσωπευτικές λανθάνουσες αναπαραστάσεις μικρότερων διαστάσεων

για τις εικόνες εισόδου, οι οποίες και στη συνέχεια αποτελούν τη βάση της όλης επεξεργασίας.

Το δίκτυο του autoencoder αποτελείται από δύο επιμέρους δίκτυα, αυτό του κωδικοποιητή (*Encoder*)  $\mathcal{E}$  και αυτό του αποκωδικοποιητή (*Decoder*)  $\mathcal{D}$ . Έτσι, δοθείσης μίας εικόνας  $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$  στον RGB χρωματικό χώρο, ο κωδικοποιητής  $\mathcal{E}$  κωδικοποιεί την εικόνα  $\mathbf{x}$  σε μία λανθάνουσα αναπαράσταση  $\mathbf{z} = \mathcal{E}(\mathbf{x})$ , ενώ ο αποκωδικοποιητής  $\mathcal{D}$  ανακατασκευάζει την αρχική εικόνα από την λανθάνουσα αυτή αναπαράσταση, δηλαδή  $\tilde{\mathbf{x}} = \mathcal{D}(\mathbf{z}) = \mathcal{D}(\mathcal{E}(\mathbf{x}))$ , όπου  $\mathbf{z} \in \mathbb{R}^{h \times w \times c}$ . Προφανώς, ο κωδικοποιητής οδηγεί σε λανθάνουσες αναπαραστάσεις μικρότερης διάστασης από την αρχική διάσταση των δεδομένων, υποδειγματοληπτώντας έτσι τις εικόνες κατά ένα παράγοντα  $f = H/h = W/w$ . Στην περίπτωση του Stable Diffusion, ο παράγοντας αυτός ισούται με 8.

Για την εκπαίδευση του δικτύου του autoencoder χρησιμοποιείται μία αντικειμενική συνάρτηση, η οποία συνδυάζει όρους perceptual (*perceptual loss*) και adversarial (*patch-based adversarial objective*) [25, 29]. Πιο συγκεκριμένα, το δίκτυο του autoencoder εκπαιδεύεται σύμφωνα με μία ανταγωνιστική (adversarial) διαδικασία [29], κατά την οποία ένας patch-based discriminator  $\mathbf{D}_\psi$  βελτιστοποιείται ώστε να διακρίνει πραγματικές εικόνες  $\mathbf{x}$  από τις αντίστοιχες ανακατασκευές  $\mathcal{D}(\mathcal{E}(\mathbf{x}))$ . Ο patch-based discriminator [25] είναι σχεδιασμένος έτσι ώστε να ανιχνεύει τη δομική γνησιότητα σε επίπεδο patches. Αυτό σημαίνει ότι προσπαθεί να ταξινομήσει καθένα από τα  $N \times N$  patches μίας εικόνας ανάλογα με το εάν είναι πραγματικά (real) ή τεχνητά (fake). Ο discriminator εφαρμόζεται με συνελκτικό τρόπο σε όλη την εικόνα και έπειτα λαμβάνεται ο μέσος όρος των επιμέρους προβλέψεων ανά patch, έτσι ώστε να εξαχθεί ένα ολικό συμπέρασμα για τη γνησιότητα της εικόνας.

Επιπλέον, προκειμένου να διασφαλιστεί η κανονικότητα στο χώρο των λανθανουσών αναπαραστάσεων (*latent space*), χρησιμοποιείται κατά την εκπαίδευση του autoencoder και ένας όρος κανονικοποίησης  $L_{reg}$ , ο οποίος εξασφαλίζει ότι η λανθάνουσα αναπαράσταση  $\mathbf{z}$  ακολουθεί μία κατανομή, η οποία παρουσιάζει μηδενική μέση τιμή και σχετικά μικρή, φραγμένη διασπορά. Στην πρωτότυπη υλοποίηση των *Rombach et al.* [22] χρησιμοποιούνται δύο διαφορετικές μέθοδοι κανονικοποίησης:

- (i) η απόσταση Kullback-Leibler μεταξύ της κατανομής του χώρου λανθανουσών αναπαραστάσεων  $q_{\mathcal{E}}(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mathcal{E}}_\mu, \boldsymbol{\mathcal{E}}_{\sigma^2})$  και της κανονικής κατανομής  $\mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$ , όπως στην περίπτωση των απλών variational autoencoders [30] και
- (ii) η χρήση ενός στρώματος κβάντωσης (*vector quantization layer*) και η μάθηση ενός codeblock  $\mathcal{Z}$ , κατάλληλων διαστάσεων [31].

Συνολικά, η αντικειμενική συνάρτηση που χρησιμοποιείται για την εκπαίδευση του autoencoder έχει την ακόλουθη μορφή:

$$L_{\text{VAE}} = \min_{\mathcal{E}, \mathcal{D}} \max_{\psi} (L_{\text{rec}}(\mathbf{x}, \mathcal{D}(\mathcal{E}(\mathbf{x}))) - L_{\text{adv}}(\mathcal{D}(\mathcal{E}(\mathbf{x}))) + \log \mathbf{D}_\psi(\mathbf{x}) + L_{\text{reg}}(\mathbf{x}; \mathcal{E}, \mathcal{D})), \quad (3.19)$$

όπου  $L_{rec}$  είναι ένας όρος ανακατασκευής, ο οποίος ελέγχει άμεσα την ποιότητα της ανακατασκευής (π.χ.  $L_1$  ή  $L_2$  μεταξύ  $\mathbf{x}$  και  $\tilde{\mathbf{x}}$ ) και ο όρος  $L_{adv}$  ισούται με,

$$L_{adv} = \log \mathcal{D}(\mathcal{E}(\mathbf{x})). \quad (3.20)$$

Στο σημείο αυτό σημειώνεται ότι στην πρωτότυπη υλοποίηση, η συνεισφορά της συνάρτησης κανονικοποίησης  $L_{reg}$  στη συνολική αντικειμενική συνάρτηση είναι πολύ μικρή και στη μεν περίπτωση (i) σταθμίζεται με ένα παράγοντα της τάξεως του  $\sim 10^{-6}$ , ενώ στην περίπτωση (ii) χρησιμοποιείται ένα codebook υψηλής διάστασης.

### 3.2.2 Denoising U-Net

Στον πυρήνα της αρχιτεκτονικής του Stable Diffusion βρίσκεται ένα Denoising Diffusion Probabilistic μοντέλο, το οποίο όπως αναφέρθηκε και στη σχετική ενότητα, εγγεί διαδοδικά θόρυβο στα διαθέσιμα δεδομένα και εν συνεχεία μέσω της αντίστροφης διαδικασίας αποθορυβοποίησης προσπαθεί να αναιρέσει την επίδραση του θορύβου, ώστε να παράξει νέα δείγματα. Μία σημαντική διαφορά σε σχέση με τα παραδοσιακά DDPMs, είναι ότι στην περίπτωση του Stable Diffusion, τόσο η διαδικασία διάχυσης όσο και η διαδικασία αποθορυβοποίησης, δεν εφαρμόζονται απευθείας στις διαθέσιμες εικόνες, αλλά στις λανθάνουσες αναπαραστάσεις αυτών, οι οποίες και λαμβάνονται μέσω του autoencoder, όπως αναλύθηκε στην ενότητα 3.2.1. Κατά αυτό τον τρόπο, μειώνεται αφενός το απαιτούμενο υπολογιστικό κόστος για την εκπαίδευση του μοντέλου διάχυσης και βελτιώνεται αφετέρου η όλη επίδοσή του, καθώς αυτό επικεντρώνεται πλέον στα σημαντικά χαρακτηριστικά των εικόνων εισόδου. Έτσι, η αντικειμενική συνάρτηση για την εκπαίδευση του μοντέλου διάχυσης της σχέσης 2.43, τροποποιείται ως εξής:

$$L_{LDM} = \mathbb{E}_{t \sim \mathcal{U}[1,T], \mathcal{E}(x), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\|\epsilon - \hat{\epsilon}_\theta(z_t, t)\|_2^2]. \quad (3.21)$$

Το νευρωνικό δίκτυο  $\hat{\epsilon}_\theta(\circ, t)$  της σχέσης 3.21 πρόκειται για ένα *time-conditional U-Net* [32]. Επομένως, κατά τη διαδικασία της εκπαίδευσης, κάθε δείγμα εισόδου  $\mathbf{x}$  μετασχηματίζεται αρχικά μέσω του autoencoder σε ένα χώρο χαμηλότερων διαστάσεων και στη συνέχεια θορυβοποιείται κατά τη διαδικασία της διάχυσης, οπότε και προκύπτει η ενθόρυβη εκδοχή του  $z_t$ . Έπειτα, η ενθόρυβη αυτή εκδοχή διέρχεται μέσω του U-Net, το οποίο προσεγγίζει το διάνυσμα θορύβου  $\hat{\epsilon}_\theta(z_t, t)$  και η ανανέωση των βαρών του γίνεται βάσει της απόκλισης της προσέγγισης αυτής από το πραγματικό διάνυσμα θορύβου  $\epsilon$ , με χρήση της μεθόδου gradient descent.

Η έως τώρα περιγραφή της λειτουργίας του μοντέλου διάχυσης αναλύει τον τρόπο με τον οποίο η βασική αρχιτεκτονική ενός DDPM προσαρμόζεται, ώστε να δέχεται ως εισόδους λανθάνουσες αναπαραστάσεις, έναντι των αρχικών δεδομένων. Δεν εξηγεί όμως τον τρόπο με τον οποίο η κειμενική περιγραφή υπεισέρχεται στην όλη διαδικασία και κατευθύνει την παραγωγή των νέων δειγμάτων. Όπως ήδη έχει αναφερθεί και στην εισαγωγή του παρόντος κεφαλαίου, βασικός σκοπός του Stable Diffusion, είναι η παραγωγή συνθετικών εικόνων, καθοδηγούμενων από αντίστοιχες κειμενικές περιγραφές. Θα πρέπει επομένως με κάποιο τρόπο, οι περιγραφές αυτές να αξιοποιούνται καταλλήλως εντός της αρχιτεκτονικής του δικτύου.

Τα μοντέλα διάχυσης, όπως και άλλοι τύποι παραγωγικών μοντέλων [23] είναι ικανά στη γενική περίπτωση, να μοντελοποιούν δεσμευμένες κατανομές. Ας υποθέσουμε ότι διαθέτουμε ένα δείγμα  $\mathbf{x}_0$ , το οποίο προέρχεται από μία δεσμευμένη κατανομή  $D(\mathbf{x}|\mathbf{y})$ , όπου  $\mathbf{y}$  μία συνθήκη, η οποία μπορεί να αναφέρεται σε κείμενο [24], σε σημασιολογικούς χάρτες [25] κτλ. Σύμφωνα με τις βασικές αρχές λειτουργίας των DDPMs, προσθέτουμε διαδοχικά Γκαουσιανό θόρυβο στο αρχικό αυτό δείγμα, βάσει της σχέσης,

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}) = \mathcal{N}(\mathbf{x}_t|\mathbf{y}; \boldsymbol{\mu}_t = \sqrt{1 - \beta_t}\mathbf{x}_{t-1}|\mathbf{y}, \boldsymbol{\Sigma}_t = \beta_t \mathbf{I}), \text{ για } t = \{0, 1, \dots, T\}. \quad (3.22)$$

Στην περίπτωση αυτή, το ενθόρυβο δείγμα  $\mathbf{x}_t$  μπορεί να εκφραστεί συναρτησί του αρχικού δείγματος  $\mathbf{x}_0$  σύμφωνα με τη σχέση,

$$q(\mathbf{x}_t|\mathbf{x}_0, \mathbf{y}) := \mathcal{N}(\mathbf{x}_t|\mathbf{y}; \sqrt{\bar{\alpha}_t}\mathbf{x}_0|\mathbf{y}, (1 - \bar{\alpha}_t) \mathbf{I}), \quad (3.23)$$

όπου ως γνωστόν  $\alpha_t = 1 - \beta_t$  και  $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$ .

Για την προσέγγιση τώρα της αντίστροφης διαδικασίας διάχυσης,

$$p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{y}) := \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_{\theta}(\mathbf{x}_t, t, \mathbf{y}), \boldsymbol{\Sigma}_{\theta}(\mathbf{x}_t, t, \mathbf{y})), \quad (3.24)$$

η οποία και αναλύθηκε στην ενότητα 3.1, χρησιμοποιείται ένα *conditional U-Net* [32], το οποίο εκπαιδεύεται βάσει της ακόλουθης αντικειμενικής συνάρτησης:

$$L_{cond} = \mathbb{E}_{t \sim \mathcal{U}[1, T], \mathbf{x}_0 \sim D(\mathbf{x}|\mathbf{y}), \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[ \left\| \boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}_{\theta} \left( \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}, t, \mathbf{y} \right) \right\|_2^2 \right]. \quad (3.25)$$

Το ερώτημα τώρα έγκειται στο πώς η συνθήκη  $\mathbf{y}$  υπεισέρχεται εντός του δικτύου αποθορυβοποίησης U-Net. Στα πλαίσια του Stable Diffusion, η συνθήκη  $\mathbf{y}$ , η οποία και πρόκειται για κειμενική περιγραφή, προεπεξεργάζεται μέσω ενός *domain-specific* encoder, ο οποίος προβάλλει την περιγραφή  $\mathbf{y}$  σε μία ενδιάμεση αναπαράσταση  $\boldsymbol{\zeta} := \boldsymbol{\tau}_{\theta}(\mathbf{y}) \in \mathbb{R}^{M \times d_{\tau}}$ . Πιο συγκεκριμένα, ο encoder αυτός υιοθετεί την αρχιτεκτονική ενός δικτύου transformer, το οποίο αποτελείται από  $N$  transformer blocks, αποτελούμενα με τη σειρά τους από global self-attention layers, layer-normalization και position-wise MLPs, όπως φαίνεται παρακάτω:

$$\begin{aligned} \boldsymbol{\zeta} &\leftarrow \text{TokEmb}(\mathbf{y}) + \text{PosEmb}(\mathbf{y}) \\ \text{for } i &= 1, \dots, N : \\ \boldsymbol{\zeta}_1 &\leftarrow \text{LayerNorm}(\boldsymbol{\zeta}) \\ \boldsymbol{\zeta}_2 &\leftarrow \text{MultiHeadSelfAttention}(\boldsymbol{\zeta}_1) + \boldsymbol{\zeta} \\ \boldsymbol{\zeta}_3 &\leftarrow \text{LayerNorm}(\boldsymbol{\zeta}_2) \\ \boldsymbol{\zeta} &\leftarrow \text{MLP}(\boldsymbol{\zeta}_3) + \boldsymbol{\zeta}_2 \\ \boldsymbol{\zeta} &\leftarrow \text{LayerNorm}(\boldsymbol{\zeta}) \end{aligned}$$

όπου για την παραγωγή της tokenized αναπαράστασης  $\text{TokEmb}(\mathbf{y})$  της κειμενικής περιγραφής  $\mathbf{y}$ , χρησιμοποιείται ένας off-the-shelf tokenizer [33].

Η παραγόμενη αναπαράσταση  $\tau_\theta(\mathbf{y})$  της κειμενικής περιγραφής προβάλλεται στα ενδιάμεσα στρώματα της αρχιτεκτονικής του δικτύου U-Net μέσω ενός στρώματος **cross-attention**, το οποίο υλοποιεί την ακόλουθη διαδικασία,

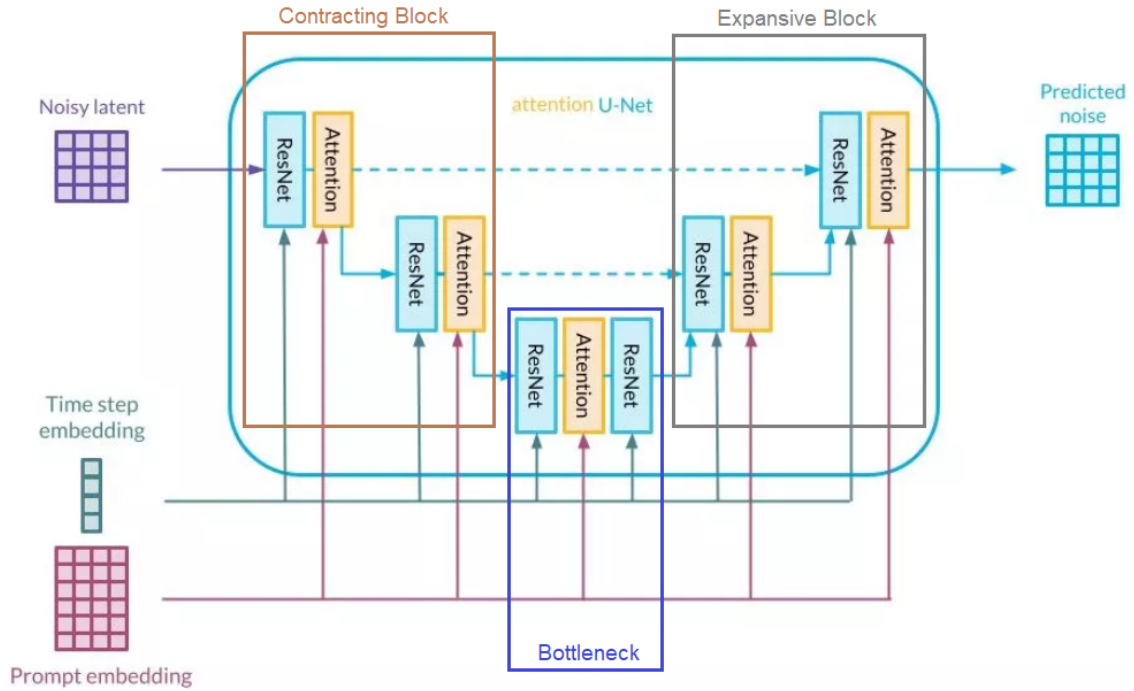
$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right) \cdot \mathbf{V}, \quad (3.26)$$

όπου

$$\mathbf{Q} = \mathbf{W}_Q^{(i)} \cdot \phi(\mathbf{z}_t), \quad \mathbf{K} = \mathbf{W}_K^{(i)} \cdot \tau_\theta(\mathbf{y}), \quad \mathbf{V} = \mathbf{W}_V^{(i)} \cdot \tau_\theta(\mathbf{y}). \quad (3.27)$$

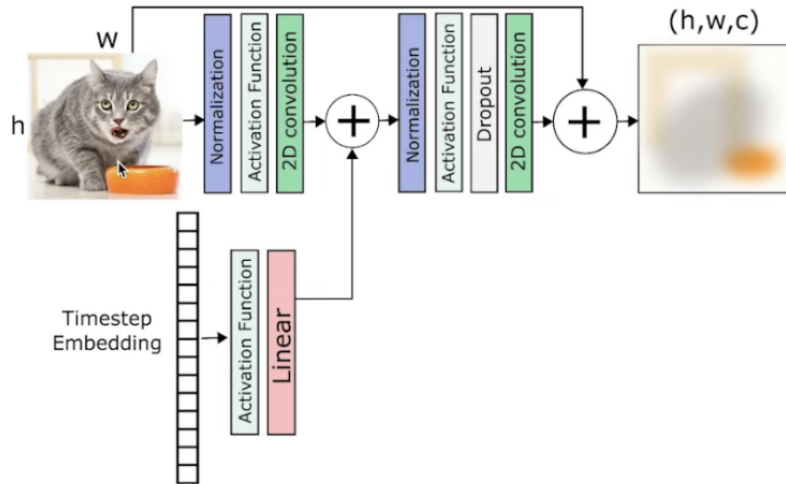
Στην παραπάνω σχέση ο όρος  $\phi(\mathbf{z}_t) \in \mathbb{R}^{N \times d_\epsilon^i}$  αναφέρεται σε μία ενδιάμεση αναπαράσταση του δικτύου U-Net, ενώ  $\mathbf{W}_V^{(i)} \in \mathbb{R}^{d \times d_\epsilon^i}$ ,  $\mathbf{W}_K^{(i)} \in \mathbb{R}^{d \times d_\tau}$ ,  $\mathbf{W}_Q^{(i)} \in \mathbb{R}^{d \times d_\tau}$  είναι τα προς μάθηση μητρώα βαρών των values, keys και queries αντίστοιχα.

Η συνολική αρχιτεκτονική του δικτύου attention U-Net που χρησιμοποιείται για την υλοποίηση της διαδικασίας αποθορυβοποίησης στο μοντέλο DDPM, παρουσιάζεται στο Σχ.3.2. Τόσο το contracting όσο και expansive block του δικτύου αποτελούνται από μία σειρά από επίπεδα, καθένα εκ των οποίων αποτελείται με τη σειρά του από ένα ResNet και ένα Attention block, με το μεν contracting block να χρησιμοποιεί 2-διάστατα convolutions για την υποδειγματοληψία, το δε expansive block 2-διάστατα deconvolutions για την υπερδειγματοληψία των εισόδων. Σε κάθε επίπεδο των διαδικασιών contracting και expansion, υπεισέρχονται στα κατάλληλα blocks, η αναπαράσταση  $\tau_\theta(\mathbf{y})$  της κειμενικής περιγραφής (prompt embedding) και μία αναπαράσταση σχετικά με το τρέχον βήμα αποθορυβοποίησης (time step embedding).



Σχήμα 3.2: Αρχιτεκτονική δικτύου attention U-Net.

Το ResNet block αναφέρεται στο βασικό residual κύτταρο των αρχιτεκτονικών *ResNet* και παρουσιάζεται στο Σχ.3.3. Στο σχήμα αυτό φαίνεται και ο τρόπος με τον οποίο η χρονική αναπαράσταση του τρέχοντος βήματος αναπαράστασης αξιοποιείται από τα επιμέρους επίπεδα του U-Net κατά τη διαδικασία της αποθορυβοποίησης.



**Σχήμα 3.3:** Δομή ResNet block της αρχιτεκτονικής του δικτύου U-Net του Σχ.3.2.

Όσον αφορά τώρα στο attention block, στην πρωτότυπη υλοποίηση των *Rombach et al.*, αυτό πρόκειται ουσιαστικά για έναν spatial transformer, με την αρχιτεκτονική που παρουσιάζεται στον Πίνακα 3.1. Στον πίνακα αυτό με  $n_h$  συμβολίζεται το πλήθος των attention κεφαλών, ενώ με  $d$  η διαστατικότητα τους.

**Πίνακας 3.1:** Αρχιτεκτονική attention block του δικτύου U-Net του Σχ.3.2.

| Layer          | Διαστάσεις Εξόδου               |
|----------------|---------------------------------|
| LayerNorm      | $h \times w \times c$           |
| Conv1 × 1      | $h \times w \times d \cdot n_h$ |
| Reshape        | $h \cdot w \times d \cdot n_h$  |
| ×T {           |                                 |
| SelfAttention  | $h \cdot w \times d \cdot n_h$  |
| MLP            | $h \cdot w \times d \cdot n_h$  |
| CrossAttention | $h \cdot w \times d \cdot n_h$  |
| Reshape        | $h \times w \times d \cdot n_h$ |
| Conv1 × 1      | $h \cdot w \times c$            |

Βάσει τώρα της ανωτέρω περιγραφής, η αντικειμενική συνάρτηση της σχέσης 3.21 που χρησιμοποιείται για την εκπαίδευση του conditional U-Net, τροποποιείται καταλλήλως ώστε να περιλαμβάνει και τη συνθήκη της κειμενικής περιγραφής, οπότε και ισούται με,

$$L_{LDM}^{cond} = \mathbb{E}_{t \sim \mathcal{U}[1, T], \mathcal{E}(x), y, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} [\|\epsilon - \hat{\epsilon}_\theta(z_t, t, \tau_\theta(y))\|_2^2]. \quad (3.28)$$

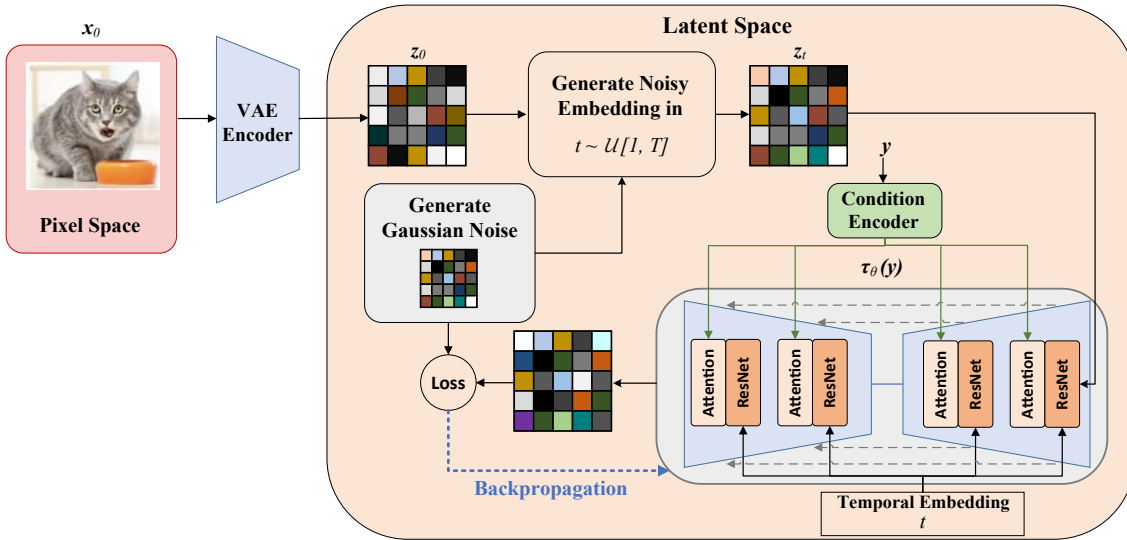
Συνολικά, η διαδικασία εκπαίδευσης του δικτύου αποθουρβοποίησης U-Net υλοποιείται σύμφωνα με τον Αλγόριθμο 3.4, ενώ παρουσιάζεται σχηματικά στο Σχ.3.4.

---

#### Αλγόριθμος 3.4: U-Net Training

---

- 1: **repeat**
  - 2:  $\mathbf{x}_0 \sim q(\mathbf{x}_0), \mathbf{y}$
  - 3:  $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0), \tau_\theta(\mathbf{y})$
  - 4:  $t \sim \text{Uniform}(\{1, \dots, T\})$
  - 5:  $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  - 6: Take gradient descent step on
 
$$\nabla_\theta \|\epsilon - \hat{\epsilon}_\theta(\sqrt{\bar{\alpha}_t} \mathbf{z}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t, \tau_\theta(\mathbf{y}))\|_2^2$$
  - 7: **until** converged
- 



Σχήμα 3.4: Διαδικασία εκπαίδευσης conditional U-Net.



## Κεφάλαιο 4

### Sketch-Guided Image Synthesis

---

Στις προηγούμενες ενότητες έγινε αναφορά στο πρόβλημα της καθοδηγούμενης από κειμενικές περιγραφές σύνθεσης εικόνων και αναλύθηκε ο τρόπος με τον οποίο το μοντέλο Stable Diffusion αντιμετωπίζει το πρόβλημα αυτό. Μία επέκταση του θεμελιώδους αυτού προβλήματος, είναι η καθοδηγούμενη από σκίτσα και κειμενικές περιγραφές σύνθεση εικόνων, κατά την οποία, πέραν της κειμενικής περιγραφής, απαιτείται και η χωρική καθοδήγηση των περιγραμμάτων των παραγόμενων εικόνων, βάσει κατάλληλων σκίτσων αναφοράς.

Οι πρωταρχικές μέθοδοι που αναπτύχθηκαν για την αντιμετώπιση του εν λόγω προβλήματος, βασίζονταν κυρίως σε παραδοσιακές τεχνικές επεξεργασίας εικόνας (π.χ. BoW, descriptors, ανίχνευση ακμών) για την ανάκτηση εικόνων από σκίτσα αναφοράς [34, 35, 36, 37, 38, 39, 40]. Οι μέθοδοι αυτές, αν και πολλά υποσχόμενες, παρουσιάζουν αρκετές αδυναμίες, όπως η περιορισμένη εκφραστικότητα, η δυσκολία χειρισμού πολύπλοκων σκίτσων και η έλλειψη φωτορεαλισμού.

Με την άνοδο της επιστήμης της βαθιάς μάθησης, οι ερευνητές άρχισαν να διερευνούν νέες προσεγγίσεις για να αντιμετωπίσουν το πρόβλημα της καθοδηγούμενης από σκίτσα σύνθεσης εικόνων, οι οποίες βασίζονται κυρίως στη χρήση GANs. Με την ανάπτυξη των υπό συνθήκη ανταγωνιστικών δικτύων, οι *Mirza et al.* [41] άνοιξαν το δρόμο για την υλοποίηση πληθώρας προβλημάτων image-to-image translation. Στο πλαίσιο αυτό, οι *Isola et al.* [25] υλοποίησαν το *pix2pix*, ένα υπό συνθήκη GAN ικανό να μαθαίνει αντιστοιχίσεις από το πεδίο της εικόνας εισόδου σε ένα επιθυμητό πεδίο εξόδου, βάσει κατάλληλου συνόλου *paired* δεδομένων. Η πρώτη ολοκληρωμένη προσπάθεια για την καθοδήγηση της διαδικασίας της σύνθεσης από σκίτσα, πραγματοποιήθηκε από τους *Chen et al.* [1], με το δίκτυό τους *SketchyGAN*. Η προτεινόμενη αρχιτεκτονική τους αποτελείται από έναν generator, με αρχιτεκτονική encoder-decoder και έναν discriminator. Η εκπαίδευση του δικτύου πραγματοποιείται με adversarial τρόπο και, εκτός από την adversarial συνάρτηση απώλειας, συμπεριλαμβάνονται και perceptual όροι απώλειας, έτσι ώστε να διασφαλιστεί ότι οι παραγόμενες εικόνες ακολουθούν οπτικά τις αντίστοιχες πραγματικές. Κατά τα επόμενα χρόνια, προτάθηκαν και αρκετές άλλες μέθοδοι βασισμένες σε GANs [3, 42, 43, 44, 45, 46], με ορισμένες εξ αυτών να εστιάζουν στην καθοδηγούμενη σύνθεση ανθρώπινων προσώπων, βάσει πορτρέτων υπό τη μορφή σκίτσων [47, 48].

Συνολικά, οι προαναφερθείσες μέθοδοι, οι οποίες βασίζονται στη χρήση GANs, αποτέλεσαν τον πρώτο αποτελεσματικό τρόπο για την αντιμετώπιση του προβλήματος της καθοδηγούμενης από σκίτσα σύνθεσης εικόνων. Παρά ταύτα, δεν παρέχουν τη δυνατότητα κειμενικής καθοδήγησης, έτσι ώστε, συγχρόνως με την ακολουθία του περιγράμματος των σκίτσων εισόδου, οι παραγόμενες εικόνες να συμμορφώνονται με αντίστοιχες κειμενικές περιγραφές. Η κειμενική αυτή απουσία, σε συνδυασμό με τα φαινόμενο του mode collapse [49] και τα προβλήματα αστάθειας, τα οποία παρουσιάζονται λόγω του ανταγωνιστικού τρόπου εκπαίδευσης των GANs, ώθησε την έρευνα στη διερεύνηση νέων μεθόδων, βασισμένων σε μοντέλα διάχυσης. Τα μοντέλα αυτά, όντας likelihood-based στον πυρήνα τους, είναι ικανά να ενσωματώνουν απρόσκοπτα στη διαδικασία της σύνθεσης την κειμενική περιγραφή, όπως άλλωστε αναλύθηκε και στο προηγούμενο κεφάλαιο, αμβλύνοντας κατά αυτό τον τρόπο τις αδυναμίες που σχετίζονται με τα GANs.

Προς την κατεύθυνση αυτή, οι *Voynov et al.* [4] προτείνουν μία γενική προσέγγιση για την καθοδήγηση της διαδικασίας παραγωγής ενός προεκπαιδευμένου text-to-image μοντέλου διάχυσης, μέσω χωρικών χαρτών. Βασική ιδέα της προσέγγισης αυτής είναι η χρήση ενός μικρού δικτύου multi-layer perceptron (MLP), το οποίο εκπαιδεύεται ώστε να προβάλλει χαρακτηριστικά των λανθανουσών αναπαραστάσεων των ενθόρυβων εικόνων σε κατάλληλους χωρικούς χάρτες. Τα χαρακτηριστικά αυτά εξάγονται από μέρη της αρχιτεκτονικής του βασικού προεκπαιδευμένου μοντέλου διάχυσης. Το εκπαιδευμένο MLP λειτουργεί εν είδει πρόβλεψης του χωρικού χάρτη των λανθανουσών αναπαραστάσεων των ενδιάμεσων παραγόμενων εικόνων και εκπαιδεύεται με ένα self-supervised τρόπο, με τη συνάρτηση απώλειας να υπολογίζεται μεταξύ του ενδιάμεσου προβλεπόμενου χωρικού χάρτη και του χάρτη στόχου. Οι *Wang et al.* [50], εκπαιδεύουν ένα μοντέλο διάχυσης χρησιμοποιώντας μια υβριδική αντικειμενική συνάρτηση. Η συνάρτηση αυτή αποτελείται από έναν identity όρο απώλειας, ο οποίος υπολογίζεται μεταξύ της εικόνας εισόδου και της ανακατασκευασμένης εκδοχής της, έπειτα από τη διαδικασία αποθορυβοποίησης και από έναν perceptual όρο απώλειας, ο οποίος ορίζεται μεταξύ του σκίτσου εισόδου και αυτού που εξάγεται από την ανακατασκευασμένη εικόνα. Ένας σημαντικότερος περιορισμός της προσέγγισής τους είναι η ανάγκη για πλήρη επανεκπαίδευση του χρησιμοποιούμενου μοντέλου διάχυσης, η οποία αποδεικνύεται ιδιαίτερος χρονοβόρα και απαιτητική ως προς τις απαιτήσεις του υλικού.

Στις επόμενες ενότητες θα γίνει περαιτέρω ανάλυση της δομής και του θεωρητικού υποβάθρου της προτεινόμενης μεθόδου των *Voynov et al.*, προκειμένου να γίνει κατανοητός ο τρόπος λειτουργίας της και θα προταθεί κατάλληλη τροποποίησή της, ώστε να βελτιωθεί η ποιότητα των παραγόμενων αποτελεσμάτων και να μειωθεί το υπολογιστικό και χρονικό κόστος.

## 4.1 Latent Edge Predictor

Όπως αναφέρθηκε και παραπάνω, στον πυρήνα της προσέγγισης του προβλήματος της καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων, σύμφωνα με τους *Voynov et al.* [4], βρίσκεται ένα δίκτυο MLP, το οποίο και εκπαιδεύεται ώστε να παράγει χωρικούς χάρτες από τις λανθάνουσες αναπαραστάσεις των ενδιάμεσων στρωμάτων της αρχιτεκτονικής του μοντέλου διάχυσης. Στα πλαίσια της υλοποίησής τους, το μοντέλο διάχυσης που χρησιμοποιείται και το οποίο αποτελεί τη βάση της όλης αρχιτεκτονικής, δεν είναι άλλο από το μοντέλο Stable Diffusion, το οποίο και αναλύεται στην ενότητα 3.2.

Βάσει λοιπόν του μοντέλου αυτού και ακολουθώντας τη μέθοδο των *Baranchuk et al.* [51], γίνεται εξαγωγή των αποτελεσμάτων από έναν καθορισμένο αριθμό ενδιάμεσων επιπέδων του δικτύου αποθρομβοποίησης U-Net (ενότητα 3.2.2), το οποίο από τούδε και στο εξής θα συμβολίζεται με  $U$ . Πιο συγκεκριμένα, για ένα tensor εισόδου  $w$ , ο οποίος συνοδεύεται από μία κειμενική περιγραφή  $y$  θα συμβολίζεται με,

$$F(w|t, \tau_\theta(y)) = [I_1(w|t, \tau_\theta(y)), \dots, I_n(w|t, \tau_\theta(y))], \quad (4.1)$$

τις concatenated ενεργοποιήσεις επιλεγμένων εσωτερικών στρωμάτων  $\{I_1, \dots, I_n\}$  του δικτύου  $U$ , όταν το  $w$  επεξεργάζεται από το δίκτυο Stable Diffusion σε επίπεδο θορύβου  $t$ . Εφόσον οι ενεργοποιήσεις διαφορετικών επιπέδων του δικτύου U-Net ενδέχεται να παρουσιάζουν διαφορετικές διαστάσεις, πριν τη συνένωσή τους, η οποία και συντελείται κατά μήκος της διάστασης των καναλιών, γίνεται αλλαγή των διαστάσεων αυτών, ώστε να συμφωνούν μεταξύ τους. Η διάσταση της εισόδου του MLP στην περίπτωση αυτή θα ισούται με το άθροισμα του αριθμού των καναλιών των επιμέρους ενεργοποιήσεων.

Το σύνολο των δεδομένων εκπαίδευσης  $\mathcal{D}$  του MLP αποτελείται από τριπλέτες της μορφής  $(x, e, y)$ , όπου  $x$  μία εικόνα,  $y$  η κειμενική περιγραφή και  $e$  ο αντίστοιχος χάρτης ακμών (*edge map*), ο οποίος ουσιαστικά αντιστοιχεί σε κάποιο ελεύθερο σκίτσο. Καθώς, όπως αναφέρθηκε και νωρίτερα, η όλη υλοποίηση βασίζεται στο προεκπαιδευμένο δίκτυο Stable Diffusion, για την κωδικοποίηση τόσο των εικόνων όσο και των χαρτών των ακμών, χρησιμοποιείται ο κωδικοποιητής  $\mathcal{E}$  του variational autoencoder του δικτύου αυτού (ενότητα 3.2.1). Προκειμένου να καταστεί εφικτή η κωδικοποίηση του χάρτη ακμών, δεδομένου ότι αυτός δίνεται σε grayscale, γίνεται αρχικά μετατροπή του σε εικόνα τριών καναλιών, μέσω της αντιγραφής του καναλιού της χρωματικής έντασης (*intensity*). Στην πράξη, το διάνυσμα εισόδου του δικτύου U-Net πρόκειται για μία ενθόρυβη εκδοχή της λανθάνουσας αναπαραστάσεως της εικόνας εισόδου  $x$ , δηλαδή,

$$z_t = \sqrt{\bar{\alpha}_t} \mathcal{E}(x) + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad (4.2)$$

όπου  $\epsilon \sim \mathcal{N}(\mathbf{0}, I)$  τυχαίος θόρυβος. Βάσει αυτού, το MLP εκπαιδεύεται ώστε να προβάλει τα concatenated χαρακτηριστικά  $F(z_t|y, t)$  στον κωδικοποιημένο χάρτη ακμών  $\mathcal{E}(e)$ .

Προκειμένου το MLP να λάβει υπόψη το επίπεδο θορύβου που χρησιμοποιείται για τη θρομβοποίηση της εικόνας εισόδου, δέχεται ως επιπλέον είσοδο, πέραν του ελεύθερου σκίτσου, το επίπεδο θορύβου  $t$  και το αντίστοιχο positional encoding αυτού, σύμφωνα με

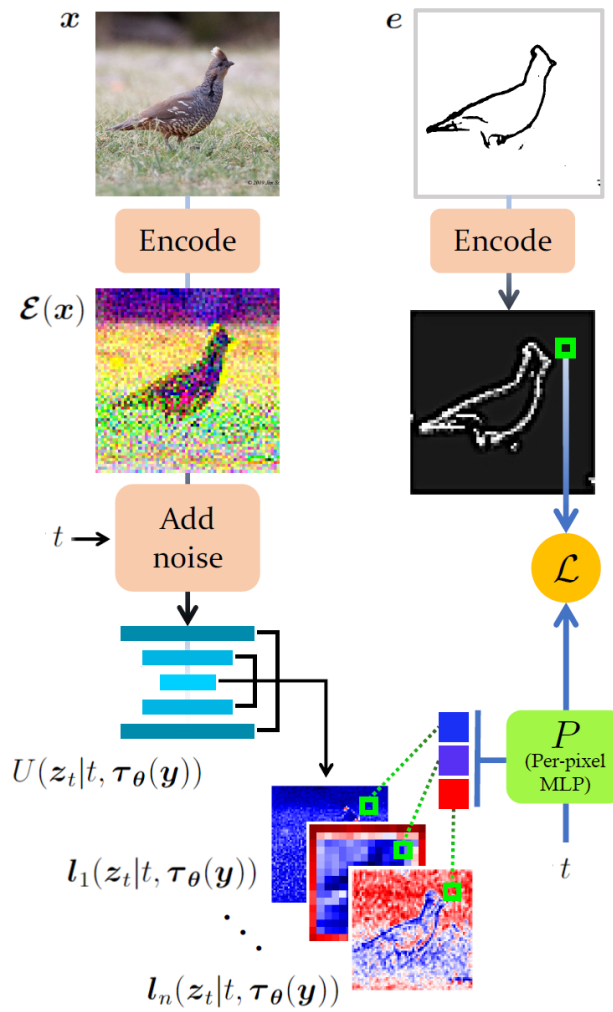
τη σχέση  $\sin(2\pi t \cdot 2^{-i})$ ,  $i = 0, \dots, 9$ , όπου  $t$  το διάνυσμα του επιπέδου θορύβου. Επομένως, το διάνυσμα  $F(\mathbf{w}|t, \tau_\theta(\mathbf{y}))$  της εξίσωσης (4.1) ισούται πρακτικά με,

$$F(\mathbf{w}|t, \tau_\theta(\mathbf{y})) = [l_1(\mathbf{w}|t, \tau_\theta(\mathbf{y})), \dots, l_n(\mathbf{w}|t, \tau_\theta(\mathbf{y})), t, \sin(2\pi t), \dots, \sin(2\pi t \cdot 2^{-9})] \quad (4.3)$$

Η διάσταση εξόδου του MLP ισούται με τον αριθμό των καναλιών εξόδου του κωδικοποιητή  $\mathcal{E}$ , η οποία στην περίπτωση του Stable Diffusion είναι ίση με  $c = 4$ . Κάθε χωρική θέση  $(i, j)$  ενός pixel στη λανθάνουσα αναπαράσταση  $F(z_t|\tau_\theta(\mathbf{y}), t)_{ij}$  μετασχηματίζεται στην αντίστοιχη θέση στο χάρτη ακμών  $\mathcal{E}(e)_{ij}$  μέσω του *per-pixel MLP* δικτύου  $P$  και ως εκ τούτου, το δίκτυο αυτό εκπαιδεύεται βάσει της ακόλουθης αντικειμενικής συνάρτησης:

$$L_{MLP} = \mathbb{E}_{(x,e,y) \sim \mathcal{D}} \mathbb{E}_{t \sim \mathcal{U}[1,T], \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \sum_{i,j} \|P(F(z_t|t, \tau_\theta(\mathbf{y})))_{ij} - \mathcal{E}(e)_{ij}\|^2, \quad (4.4)$$

όπου το δίκτυο  $P$  εφαρμόζεται σε κάθε latent pixel ανεξάρτητα. Στο Σχ.4.1 παρουσιάζεται συνολικά η διαδικασία εκπαίδευσης του per-pixel MLP.



Σχήμα 4.1: Σχήμα εκπαίδευσης per-pixel MLP latent edge predictor [4].

Προφανώς, προκειμένου να καταστεί εφικτή η σύγκριση του εξαγόμενου χάρτη ακμών  $\mathbf{P}(\mathbf{F}(z_t|t, \tau_\theta(\mathbf{y})))$  με την αντίστοιχη λανθάνουσα αναπαράσταση του σκίτσου αναφοράς, θα πρέπει οι διαστάσεις των δύο αυτών μητρώων να συμφωνούν. Όπως αναφέρθηκε και στην ενότητα 3.2.1, η λανθάνουσα αναπαράσταση  $\mathcal{E}(\mathbf{e})$  του χάρτη ακμών  $\mathbf{e}$  είναι διαστάσεων  $[W/f, H/h, c]$ , όπου  $W, H$  το πλάτος και το ύψος της εικόνας εισόδου, δηλαδή του αρχικού σκίτσου,  $f = 8$  ο παράγοντας κλιμάκωσης και  $c = 4$  το πλήθος των καναλιών εξόδου του variational encoder του Stable Diffusion. Εφόσον το δίκτυο MLP δέχεται ως εισόδους διανύσματα, το μητρώο  $\mathbf{F}(z_t|t, \tau_\theta(\mathbf{y}))$  των concatenated ενεργοποιήσεων των εσωτερικών στρωμάτων του U-Net, διαστάσεων  $[W/8, H/8, c_F]$ , μετατρέπεται αρχικά στη flattened εκδοχή του, διαστάσεων  $\left[\frac{W \times H}{8^2}, c_F\right]$ , η οποία και δίνεται εν συνεχεία ως είσοδος στο δίκτυο. Τέλος, η έξοδος του MLP, η οποία και είναι διαστάσεων  $\left[\frac{W \times H}{8^2}, 4\right]$  αναδιατάσσεται, ώστε να μετατραπεί και πάλι σε ένα μητρώο διαστάσεων  $[W/8, H/8, 4]$ , το οποίο και μπορεί πλέον να συγκριθεί με τη λανθάνουσα αναπαράσταση  $\mathcal{E}(\mathbf{e})$ . Βάσει της περιγραφής αυτής καθίσταται φανερό, ότι το MLP λειτουργεί με έναν per-pixel τρόπο, αντιμετωπίζοντας μεμονωμένα το κάθε pixel της εκάστοτε εισόδου, προκειμένου να διαπιστώσει εάν αυτό ανήκει σε κάποια ακμή. Το flattening της εισόδου οδηγεί σε διάσπαση της χωρικής διάταξης των pixels και ως εκ τούτου σε απώλεια της χωρικής πληροφορίας. Κατά αυτό τον τρόπο προσδίδεται μια αμεροληψία στην όλη διαδικασία, ενώ με την αναδιάταξη της εξόδου ανακατασκευάζεται η χωρική μορφή του προβλεπόμενου χάρτη, ο οποίος και χρησιμοποιείται εν συνεχεία για τον υπολογισμό της αντικειμενικής συνάρτησης. Στη συνέχεια παρουσιάζεται συνολικά ο αλγόριθμος εκπαίδευσης του MLP, σύμφωνα με όσα αναλύθηκαν έως τώρα.

---

#### Αλγόριθμος 4.1: Per-Pixel MLP Latent Edge Predictor Training

---

**Require:** max\_epochs

- 1: **for** epoch in range(max\_epochs) **do**
  - 2:     **for** every triplet  $(\mathbf{x}, \mathbf{e}, \mathbf{y})$  in dataset **do**
  - 3:         Get the latent representation  $\mathcal{E}(\mathbf{x})$
  - 4:         Get the latent representation  $\mathcal{E}(\mathbf{e})$
  - 5:          $t \sim \text{Uniform}(\{1, \dots, T\})$
  - 6:          $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  - 7:         Add noise to  $\mathcal{E}(\mathbf{x})$  and get the noisy representation
 
$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t} \mathcal{E}(\mathbf{x}) + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon},$$
  - 8:         Pass the noisy representation  $\mathbf{z}_t$  through the U-Net
  - 9:         Extract the intermediate activations  $\mathbf{F}(z_t|t, \tau_\theta(\mathbf{y}))$
  - 10:         Pass  $\mathbf{F}(w|t, \tau_\theta(\mathbf{y}))$  through the per-pixel MLP
  - 11:         Take gradient step on  $L_{MLP}$  (eq.4.4)
  - 12:     **end for**
  - 13: **end for**
-

Όταν ολοκληρωθεί η εκπαίδευση μέσω της αντικειμενικής συνάρτησης της σχέσης 4.4, το μοντέλο  $P$  λειτουργεί πλέον ως ένας *per-spatial location differential predictor* κωδικοποιημένων ακμών, για κωδικοποιημένες εικόνες με προστιθέμενο θόρυβο σε επίπεδο  $t$ . Καθώς το δίκτυο αυτό εφαρμόζεται per-pixel, η εκπαίδευσή του γίνεται υπό έναν τοπικό τρόπο, με αποτέλεσμα το μοντέλο να είναι αγνωστικό (*agnostic*) ως προς το domain της εικόνας. Τέλος, η per-pixel αυτή λειτουργία, επιτρέπει τη χρήση ενός σχετικά μικρού συνόλου δεδομένων (μερικές χιλιάδες εικόνες) για την εκπαίδευση του μοντέλου.

### Αρχιτεκτονική δικτύου MLP

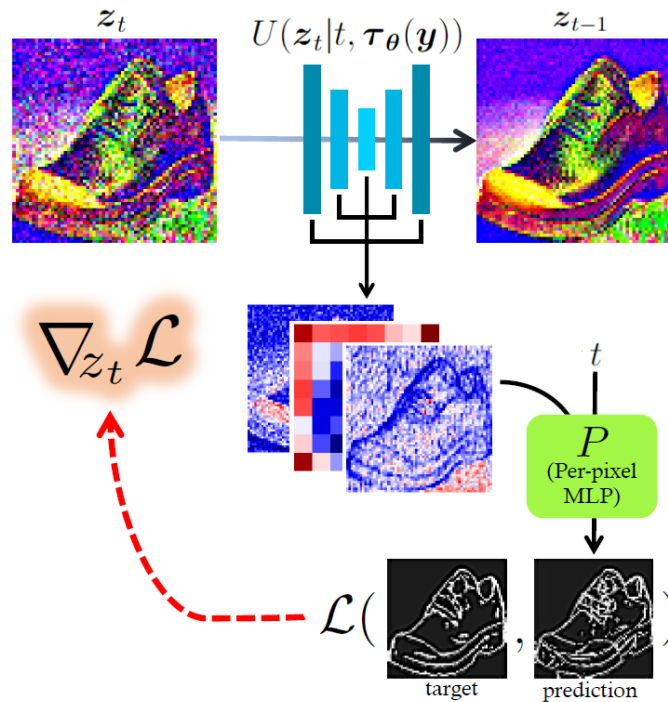
Το per-pixel MLP που αναλύθηκε στην προηγούμενη ενότητα, πρόκειται για ένα απλό feedforward πολυστρωματικό δίκτυο perceptron, το οποίο στην αρχική υλοποίηση των *Voynov et al.* [4] αποτελείται από 4 πλήρως συνδεδεμένα κρυφά στρώματα νευρώνων, διαστάσεων 512, 256, 128, 64 αντίστοιχα, με συναρτήσεις ενεργοποίησης ReLU, ακολουθούμενα από στρώματα batch normalization. Η συνολική αρχιτεκτονική του εν λόγω δικτύου παρουσιάζεται στον Πίνακα 4.1.

**Πίνακας 4.1:** Αρχιτεκτονική του per-pixel MLP [4].

|    | Layers  | Output Shape |
|----|---|--------------|
| 1. | $\begin{bmatrix} \text{Linear} \\ \text{ReLU} \\ \text{BatchNormalization} \end{bmatrix}$ | [-1, 512]    |
| 2. | $\begin{bmatrix} \text{Linear} \\ \text{ReLU} \\ \text{BatchNormalization} \end{bmatrix}$ | [-1, 256]    |
| 3. | $\begin{bmatrix} \text{Linear} \\ \text{ReLU} \\ \text{BatchNormalization} \end{bmatrix}$ | [-1, 128]    |
| 4. | $\begin{bmatrix} \text{Linear} \\ \text{ReLU} \\ \text{BatchNormalization} \end{bmatrix}$ | [-1, 64]     |
| 5. | Linear  | [-1, 4]      |

## 4.2 Παραγωγή νέων εικόνων

Σκοπός της καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης είναι, δοθέντος ενός σκίτσου  $e$  και μίας κειμενικής περιγραφής  $y$ , η παραγωγή μίας υψηλής ποιότητας εικόνας, η οποία ακολουθεί τόσο την κειμενική περιγραφή, όσο και το βασικό περίγραμμα του σκίτσου. Στο πλαίσιο αυτό, οι *Voynov et al.* [4] προτείνουν τη μέθοδο που παρουσιάζεται στο Σχ.4.2.



Σχήμα 4.2: Sketch-Guided Text-to-Image Synthesis [4].

Η διαδικασία της παραγωγής ξεκινά με τη δειγματοληψία τυχαίου Γκαουσιανού θορύβου  $x_T$  και την παραγωγή της αντίστοιχης λανθάνουσας αναπαράστασης  $z_T$ , μέσω του κωδικοποιητή του δικτύου του Stable Diffusion, δηλαδή  $z_T = \mathcal{E}(x_T)$ . Στη συνέχεια, στην περίπτωση των DDPMs, η διαδικασία της αποθορυβοποίησης περιλαμβάνει στη γενική περίπτωση  $T$  διαδοχικά βήματα  $z_t \rightarrow z_{t-1}$  έως ότου παραχθεί το δείγμα  $z_0$ , το οποίο και αποτελεί τη λανθάνουσα αναπαράσταση της τελικής εξόδου  $x_0$ . Βάσει της λογικής αυτής, σε κάθε ενδιάμεσο βήμα  $t$  υλοποιούνται τα εξής:

- Γίνεται εκτίμηση του διανύσματος θορύβου  $\hat{\epsilon}_\theta(z_t, \tau_\theta(y), t)$  και βάσει αυτού και κάποιου αλγόριθμου δειγματοληψίας υπολογίζεται το επόμενο δείγμα  $z_{t-1}$ .
- Εξάγεται το διάνυσμα των concatenated ενεργοποιήσεων των ενδιάμεσων στρωμάτων του U-Net  $F(z_t|t, \tau_\theta(y))$ , σύμφωνα με τη σχέση 4.1.
- Το εκπαιδευμένο πλέον per-pixel MLP  $P$  της ενότητας 4.1 παράγει βάσει του διανύσματος  $F(z_t|t, \tau_\theta(y))$  τον προβλεπόμενο χάρτη ακμών  $P(F(z_t|t, \tau_\theta(y)))$ , ο οποίος εν συνεχεία συγκρίνεται με τη λανθάνουσα αναπαράσταση του δοθέντος χάρτη ακμών  $\mathcal{E}(e)$ ,

$$\mathcal{L}(\mathbf{P}(\mathbf{F}(z_t|t, \tau_\theta(\mathbf{y}))), \mathcal{E}(\mathbf{e})) = \|\mathbf{P}(\mathbf{F}(z_t|t, \tau_\theta(\mathbf{y}))) - \mathcal{E}(\mathbf{e})\|^2. \quad (4.5)$$

- Ακολουθώντας το παράδειγμα των *Dhariwal* και *Nichol* [27], υπολογίζεται η ποσότητα  $-\nabla_{z_t} \mathcal{L}$ , η οποία είναι και αυτή που καθοδηγεί τη σύνθεση βάσει του σκίτσου. Διαισθητικά, αυτή η αντιπαράγωγος ωθεί ένα ενδιάμεσο δείγμα  $z_t$  ώστε να ακολουθεί κατά το δυνατόν το περίγραμμα του δοθέντος σκίτσου. Θεωρώντας λοιπόν, ότι το προβλεπόμενο επόμενο δείγμα είναι το  $z_{t-1}$ , λαμβάνοντας υπόψιν τον προαναφερθέν όρο της αντιπαράγωγου, το νέο, διορθωμένο προβλεπόμενο δείγμα δίνεται από τη σχέση,

$$z_{t-1} \leftarrow z_{t-1} - \alpha \nabla_{z_t} \mathcal{L}(\mathbf{P}(\mathbf{F}(z_t|t, \tau_\theta(\mathbf{y}))), \mathcal{E}(\mathbf{e})), \quad (4.6)$$

όπου ο παράγοντας  $\alpha$  ελέγχει το βαθμό στον οποίο η καθοδήγηση λόγω σκίτσου επηρεάζει το παραγόμενο δείγμα. Στην πράξη, η παράμετρος αυτή δίνεται από τη σχέση,

$$\alpha = \frac{\|z_t - z_{t-1}\|_2}{\|\nabla_{z_t} \mathcal{L}(\mathbf{P}(\mathbf{F}(z_t|t, \tau_\theta(\mathbf{y}))), \mathcal{E}(\mathbf{e}))\|_2} \cdot \beta, \quad (4.7)$$

όπου  $\beta$  μία υπερπαράμετρος, η οποία παραμένει σταθερή κατά τη διαδικασία της σύνθεσης και η οποία ορίζεται εκ των προτέρων από το χρήστη. Μεγάλες τιμές της παραμέτρου αυτής οδηγούν σε παραγόμενες εικόνες, οι οποίες ακολουθούν με μεγάλη ακρίβεια το περίγραμμα του δοθέντος σκίτσου, υστερώντας ωστόσο σε ρεαλισμό και φυσικότητα.

Προφανώς, η όλη ανωτέρω διαδικασία της καθοδηγούμενης σύνθεσης εμπεριέχει κάποιες υπερπαραμέτρους, οι οποίες ορίζονται από τον εκάστοτε χρήστη. Η πρώτη εξ αυτών, η οποία από τούδε και στο εξής θα συμβολίζεται με  $S$ , καθορίζει τον αριθμό των βημάτων της διαδικασίας αποθουροποίησης, κατά τα οποία θα εφαρμόζεται η καθοδήγηση λόγω του περιγράμματος του σκίτσου. Η παράμετρος αυτή είναι ιδιαίτερος σημαντική, με μικρές τιμές της να οδηγούν σε αδυναμία αποτύπωσης του περιγράμματος του σκίτσου αναφοράς και μεγάλες τιμές της να οδηγούν σε έντονη πιστότητα ως προς το σκίτσο αυτό, σε βάρος ωστόσο του ρεαλισμού. Η δεύτερη υπερπαράμετρος είναι η σταθερά  $\beta$  της σχέσης 4.7, η οποία όπως αναφέρθηκε και παραπάνω, ελέγχει το ποσοστό συνεισφοράς της καθοδήγησης στη διαδικασία της αποθουροποίησης. Πρακτικά, οι δύο αυτές υπερπαραμέτροι λειτουργούν συνδυαστικά και επηρεάζουν σε πολύ μεγάλο βαθμό το ρεαλισμό του τελικού παραγόμενου δείγματος, καθώς και την πιστότητά του ως προς το αντίστοιχο σκίτσο αναφοράς. Τέλος, σημαίνοντα ρόλο στην όλη διαδικασία διαδραματίζει και η μέθοδος που θα χρησιμοποιηθεί για την υλοποίηση της διαδικασίας αποθουροποίησης (Κεφάλαιο 2), και κατά συνέπεια ο αριθμός των βημάτων αποθουροποίησης που θα επιλεχθεί. Περαιτέρω μελέτη για την επίδραση των παραμέτρων αυτών στη διαδικασία της σύνθεσης γίνεται στο επόμενο μέρος της παρούσας εργασίας, όπου και παρουσιάζονται τα πειραματικά αποτελέσματα.

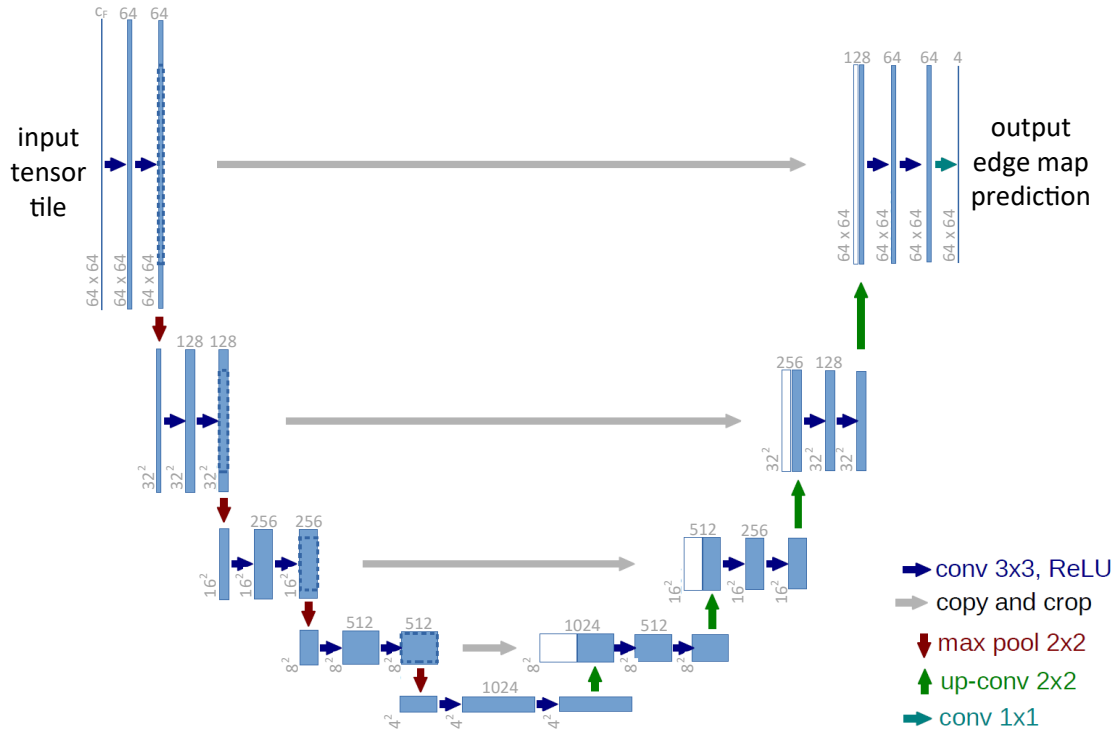


### 4.3 Τροποποίηση Αρχιτεκτονικής Latent Edge Predictor

Όπως αναφέρθηκε στην ενότητα 4.1, ως αρχιτεκτονική του latent edge predictor χρησιμοποιείται ένα πολυστρωματικό feedforward δίκτυο perceptron, το οποίο λειτουργεί με αγνωστικό τρόπο, εξετάζοντας μεμονωμένα το κάθε pixel εισόδου ως προς το εάν αυτό αποτελεί pixel ακμής. Ο αγνωστικός αυτός χαρακτήρας του μοντέλου, απεξαρτητοποιεί μεν τη λειτουργία του από το εκάστοτε domain των εικόνων εισόδου, δε λαμβάνει υπόψιν δε τυχούσες χωρικές πληροφορίες, οι οποίες μπορεί σε κάποιες περιπτώσεις να είναι ιδιαίτερος χρήσιμες και ωφέλιμες για την εξαγωγή πληροφοριών σχετικά με το είδος των pixels.

Προς την κατεύθυνση αυτή και προκειμένου η μελέτη της καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης να είναι πιο πλήρης, στα πλαίσια της παρούσας εργασίας, προτείνεται η τροποποίηση της αρχικής αρχιτεκτονικής του latent edge predictor, με τέτοιο τρόπο, ώστε αυτός να λαμβάνει πλέον υπόψιν και χωρικές πληροφορίες των εισόδων του. Για το σκόπο αυτό, το αρχικό δίκτυο MLP αντικαθίσταται από ένα συνελικτικό δίκτυο τύπου U-Net [32]. Η επιλογή του δικτύου αυτού γίνεται βάσει δύο σημαντικών ιδιοτήτων του. Η πρώτη εξ αυτών σχετίζεται με την ικανότητά του να εντοπίζει χωρικές συσχετίσεις και εξαρτήσεις των εισόδων του, όντας συνελικτικό νευρωνικό δίκτυο, ενώ η δεύτερη οφείλεται στη συμμετρική αρχιτεκτονική του, η οποία επιτρέπει την παραγωγή μητρώων επιθυμητών διαστάσεων και καναλιών στην έξοδό του. Η δεύτερη αυτή ιδιότητα είναι υψίστης σημασίας, καθώς οι εξαγόμενοι χάρτες ακμών πρέπει να είναι διαστάσεων  $[w, h, c]$ , όπου  $c$  το πλήθος των καναλιών και  $w = W/f, h = H/f$  οι διαστάσεις της εξόδου του variational encoder, με  $W, H$  το πλάτος και το ύψος των εικόνων εισόδου αντίστοιχα και  $f$  ο παράγοντας κλιμάκωσης του variational encoder. Στην περίπτωση του Stable Diffusion, το πλήθος των καναλιών εξόδου ισούται με  $c = 4$ , ενώ ο παράγοντας κλιμάκωσης με  $f = 8$ . Επομένως, οι εξαγόμενοι χάρτες ακμών του latent edge predictor πρέπει να είναι διαστάσεων  $[W/8, H/8, 4]$ , ώστε να μπορούν να συγκριθούν με άμεσο τρόπο με την αντίστοιχη λανθάνουσα αναπαράσταση του σκίτου αναφοράς. Η παραγωγή μητρώων τέτοιων διαστάσεων στην έξοδο του U-Net αποτελεί μία ευθεία διαδικασία και δεν απαιτεί κάποια μορφή flattening της εισόδου και έπειτα αναδιάταξη του τελικού αποτελέσματος, όπως στην περίπτωση του MLP.

Όσον αφορά τώρα στα τεχνικά χαρακτηριστικά, η αρχιτεκτονική του δικτύου U-Net που επιλέγεται, αποτελείται από ένα πρώτο συνελικτικό επίπεδο, το οποίο πραγματοποιεί μία αρχική εξαγωγή χαρακτηριστικών της εισόδου, το οποίο ακολουθείται από τα υποδίκτυα του encoder, του bottleneck και του decoder. Τα υποδίκτυα του encoder και του decoder παρουσιάζουν προφανώς πλήρως συμμετρική δομή και περιλαμβάνουν συνολικά 3 επαναλαμβανόμενα blocks, καθένα εκ των οποίων αποτελείται από συνελικτικά στρώματα, συναρτήσεις ενεργοποίησης ReLU και στρώματα batch normalization. Η συνολική αρχιτεκτονική του δικτύου παρουσιάζεται σχηματικά στο Σχ.4.3 και αναλυτικά στον Πίνακα 4.2. Με  $c_F$  συμβολίζεται το πλήθος των καναλιών του διανύσματος  $F$  (σχέση 4.3) των ενδιάμεσων ενεργοποιήσεων, επαυξημένο με το διάνυσμα του επιπέδου θορύβου και των αντίστοιχων positional encodings.



Σχήμα 4.3: Αρχιτεκτονική U-Net latent edge predictor.

Στο σημείο αυτό κρίνεται σκόπιμο να γίνει μία πολύ σημαντική παρατήρηση. Στα πλαίσια της παρούσας εργασίας, ως μέσο για την καθοδηγούμενη από κειμενική περιγραφή σύνθεση εικόνων χρησιμοποιείται, όπως ήδη έχει αναφερθεί το latent μοντέλο διάχυσης Stable Diffusion. Το προεκπαιδευμένο αυτό μοντέλο παράγει στη γενική περίπτωση εικόνες μεταβλητών διαστάσεων, από  $64 \times 64$ , έως και  $1024 \times 1024$ . Παρά ταύτα, τα βέλτιστα αποτελέσματα παρατηρούνται στην περίπτωση εικόνων διαστάσεων  $512 \times 512$ . Με γνώμονα την παρατήρηση αυτή, σε όλες τις υλοποιήσεις της παρούσας εργασίας, οι εικόνες εισόδου, έχουν διαστάσεις  $[W, H, 3] = [512, 512, 3]$  και ως εκ τούτου, οι αντίστοιχες λανθάνουσες αναπαραστάσεις που παράγονται από τον variational encoder έχουν διαστάσεις  $[W/8, H/8, 4] = [64, 64, 4]$ .

Έχοντας τώρα ορίσει την αρχιτεκτονική του δικτύου U-Net, το οποίο και θα αντικαταστήσει το αντίστοιχο MLP, λειτουργώντας ως latent edge predictor, θα πρέπει η διαδικασία της εκπαίδευσης να τροποποιηθεί καταλλήλως, ώστε να είναι συμβατή με τη νέα αυτή αρχιτεκτονική. Για λόγους απλότητας και ευκολίας, ο latent edge predictor ο οποίος βασίζεται στην αρχιτεκτονική του δικτύου U-Net, θα συμβολίζεται από εδώ και στο εξής με  $U_{LEP}$ . Η εκπαίδευση του δικτύου αυτού, είναι πανομοιότυπη με αυτή της περίπτωσης του δικτύου MLP (Σχ.4.1, Αλγόριθμος 4.1), με τη μόνη διαφορά ότι έπειτα από την εξαγωγή του μητρώου των ενδιάμεσων ενεργοποιήσεων  $F(z_t|t, \tau_\theta(y))$ , δεν απαιτείται κάποια περαιτέρω τροποποίησή του, όπως για παράδειγμα το flattening που συμβαίνει στην περίπτωση του MLP, οπότε και το μητρώο δίνεται απευθείας ως είσοδος στο δίκτυο, ώστε να παραχθεί ο προβλεπόμενος χάρτης ακμών  $U_{LEP}(F(z_t|t, \tau_\theta(y)))$ . Τέλος, αντίστοιχη εί-

ναι και η διαδικασία της καθοδηγούμενης παραγωγής εικόνων με χρήση του  $U_{LEP}$ , όπως αυτή περιγράφεται στην ενότητα 4.2, με μοναδική διαφορά αυτή που μόλις αναφέρθηκε.

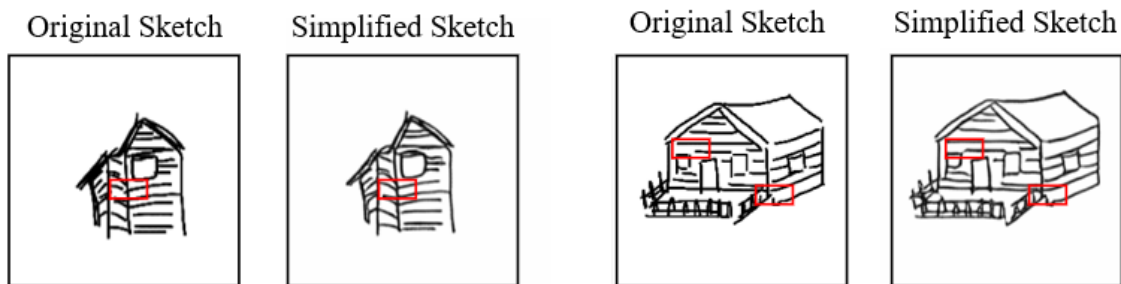
**Πίνακας 4.2:** Αρχιτεκτονική U-Net latent edge predictor. Με  $c_o, W, H$  συμβολίζονται το πλήθος των καναλιών, το πλάτος και το ύψος των μητρώων εξόδου του εκάστοτε επιπέδου του δικτύου, αντίστοιχα. Κάθε συνελκτικό στρώμα (Conv  $i, i = 1, \dots, 19$ ) ακολουθείται από ένα στρώμα batch normalization και από συνάρτηση ενεργοποίησης ReLU.

| Unit Level        | Layers   | Filter                | Stride             | Output Shape ( $[c_o, W, H]$ ) |                 |
|-------------------|----------|-----------------------|--------------------|--------------------------------|-----------------|
| <b>Input</b>      |          |                       |                    | $[c_F, 64, 64]$                |                 |
| <b>Encoder</b>    | Level 1  | Conv 1                | $3 \times 3, 64$   | 1                              | $[64, 64, 64]$  |
|                   |          | Conv 2                | $3 \times 3, 64$   | 1                              | $[64, 64, 64]$  |
|                   |          | MaxPool               |                    |                                | $[64, 32, 32]$  |
|                   | Level 2  | Conv 3                | $3 \times 3, 128$  | 1                              | $[128, 32, 32]$ |
|                   |          | Conv 4                | $3 \times 3, 128$  | 1                              | $[128, 32, 32]$ |
|                   |          | MaxPool               |                    |                                | $[128, 16, 16]$ |
|                   | Level 3  | Conv 5                | $3 \times 3, 256$  | 1                              | $[256, 16, 16]$ |
|                   |          | Conv 6                | $3 \times 3, 256$  | 1                              | $[256, 16, 16]$ |
|                   |          | MaxPool               |                    |                                | $[256, 8, 8]$   |
|                   | Level 4  | Conv 7                | $3 \times 3, 512$  | 1                              | $[512, 8, 8]$   |
|                   |          | Conv 8                | $3 \times 3, 512$  | 1                              | $[512, 8, 8]$   |
|                   |          | MaxPool               |                    |                                | $[512, 4, 4]$   |
| <b>Bottleneck</b> | Level 5  | Conv 9                | $3 \times 3, 1024$ | 1                              | $[1024, 4, 4]$  |
|                   |          | Conv 10               | $3 \times 3, 1024$ | 1                              | $[1024, 4, 4]$  |
| <b>Decoder</b>    | Level 6  | TransposedConvolution | $2 \times 2, 512$  | 2                              | $[512, 8, 8]$   |
|                   |          | Conv 11               | $3 \times 3, 512$  | 1                              | $[512, 8, 8]$   |
|                   |          | Conv 12               | $3 \times 3, 512$  | 1                              | $[512, 8, 8]$   |
|                   | Level 7  | TransposedConvolution | $2 \times 2, 256$  | 2                              | $[256, 16, 16]$ |
|                   |          | Conv 13               | $3 \times 3, 256$  | 1                              | $[256, 16, 16]$ |
|                   |          | Conv 14               | $3 \times 3, 256$  | 1                              | $[256, 16, 16]$ |
|                   | Level 8  | TransposedConvolution | $2 \times 2, 128$  | 2                              | $[128, 32, 32]$ |
|                   |          | Conv 15               | $3 \times 3, 128$  | 1                              | $[128, 32, 32]$ |
|                   |          | Conv 16               | $3 \times 3, 128$  | 1                              | $[128, 32, 32]$ |
|                   | Level 9  | TransposedConvolution | $2 \times 2, 64$   | 2                              | $[64, 64, 64]$  |
|                   |          | Conv 17               | $3 \times 3, 64$   | 1                              | $[64, 64, 64]$  |
|                   |          | Conv 18               | $3 \times 3, 64$   | 1                              | $[64, 64, 64]$  |
| <b>Output</b>     | Level 10 | Conv 19               | $3 \times 3, 4$    | 1                              | $[4, 64, 64]$   |

#### 4.4 Προσθήκη Δικτύου Απλοποίησης Σκίτσων

Έχοντας ολοκληρώσει στο σημείο αυτό την ανάλυση και την περιγραφή της διαδικασίας της καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων, κρίνεται σκόπιμο να γίνει μια πολύ σημαντική παρατήρηση. Εφόσον σκοπός της σύνθεσης είναι οι τελικές παραγόμενες εικόνες να ακολουθούν κατά το δυνατόν τα αντίστοιχα σκίτσα αναφοράς, διατηρώντας συγχρόνως το ρεαλισμό και τη φυσικότητά τους, εύκολα αντιλαμβάνεται κανείς, ότι σε περίπτωση που τα σκίτσα αυτά δεν είναι αρκούντως ευδιάκριτα, η όλη διαδικασία δυσχεραίνεται σημαντικά. Το γεγονός αυτό, σε συνδυασμό με τον εκ φύσεως ακατάστατο και πρόχειρο χαρακτήρα που παρουσιάζουν τα σκίτσα, καθιστά σημαντική την ύπαρξη κάποιου μηχανισμού, μέσω του οποίου θα επιτρέπεται η απλοποίηση και η ομαλοποίηση ορισμένων εξ αυτών, τα οποία και είναι πολύ δυσδιάκριτα και περίπλοκα.

Προς την κατεύθυνση αυτή και προκειμένου να ενισχυθεί η σθεναρότητα της όλης διαδικασίας, γίνεται χρήση ενός δικτύου απλοποίησης σκίτσων (*sketch simplification network*)  $\mathcal{S}$ , το οποίο και προστίθεται στην όλη διαδικασία, πριν από την τροφοδότηση του σκίτσου αναφοράς στον variational encoder του Stable Diffusion. Το δίκτυο αυτό [52, 53], πρόκειται για ένα συνελκτικό δίκτυο, το οποίο ακολουθεί μία συμμετρική αρχιτεκτονική και αποτελείται συνολικά από 23 επιμέρους συνελκτικά επίπεδα. Η ιδιαιτερότητα του δικτύου έγκειται στον τρόπο με τον οποίο αυτό εκπαιδεύεται. Πιο συγκεκριμένα, η εκπαίδευση του δικτύου πραγματοποιείται με έναν υβριδικό μηχανισμό, ο οποίος συνδυάζει τις τεχνικές της επιβλεπόμενης και μη επιβλεπόμενης μάθησης. Για το επιβλεπόμενο μέρος της εκπαίδευσης, αξιοποιούνται ζεύγη τα οποία αποτελούνται από σκίτσα και απλοποιημένες εκδοχές αυτών και η ανανέωση των βαρών του δικτύου γίνεται βάσει της ελαχιστοποίησης του μέσου τετραγωνικού σφάλματος των εξόδων του και των απλοποιημένων εκδοχών των σκίτσων. Κατά αυτό τον τρόπο εξασφλίζεται ότι οι εξοδοί του δικτύου θα ακολουθούν τη χωρική διάταξη των σκίτσων εισόδου. Όσον αφορά τώρα στο μέρος τη μη επιβλεπόμενης μάθησης, αυτό εμπνέεται από την οικογένεια των GANs και υλοποιείται μέσω της χρήσης ενός δικτύου discriminator. Ο discriminator αυτός λειτουργεί επικουρικά και εκπαιδεύεται ώστε να διακρίνει τις γραμμές που αντιστοιχούν σε πραγματικά σκίτσα, από εκείνες οι οποίες παράγονται από το δίκτυο απλοποίησης. Κατά αντίστοιχο τρόπο, το δίκτυο απλοποίησης εκπαιδεύεται ώστε να ξεγελάει τον discriminator, καθιστώντας τον ανίκανο να διακρίνει τα πραγματικά από τα παραγόμενα σκίτσα. Στο Σχ.7.6 παρουσιάζονται δύο παραδείγματα αρχικών και απλοποιημένων σκίτσων.



Σχήμα 4.4: Αρχικά και απλοποιημένα σκίτσα με χρήση του sketch simplification network .

Από το Σχ.7.6 προηγούμενης σελίδας διαπιστώνεται ότι η διέλευση των σκίτσων μέσω του sketch simplification network, οδηγεί σε νέα σκίτσα, τα οποία ακολουθούν ως προς τη γεωμετρία τα σκίτσα εισόδου και στα οποία παρατηρείται μια γενική ομαλοποίηση των γραμμών, εξασθένιση των πολλαπλών αλληλοεπικαλυπτόμενων γραμμών και συνένωση κάποιων πολύ κοντινών γραμμών (κόκκινα πλαίσια).

Όπως αναφέρθηκε και ανωτέρω, το δίκτυο απλοποίησης σκίτσων προστίθεται πριν το στάδιο της κωδικοποίησης του σκίτσου αναφοράς, ενώ η χρήση του επαφίεται κάθε φορά στην επιλογή του χρήστη. Ενσωματώνοντας λοιπόν και το δίκτυο αυτό, προκύπτει η τελική μορφή της διαδικασίας της καθοδηγούμενης από σκίτσα σύνθεσης εικόνων, η οποία και παρουσιάζεται συνολικά στο Σχ.4.5 της επόμενης σελίδας. Η καθοδήγηση της σύνθεσης βάσει σκίτσου παρουσιάζεται ενδεικτικά για ένα βήμα στο πρώτο μέρος του χώρου των λανθανουσών αναπαραστάσεων και έπειτα επαναλαμβάνεται για  $S$  συνολικά βήματα, όπως έχει ήδη αναφερθεί και στην ενότητα 4.2. Ακολουθεί υπό μορφή αλγορίθμου η περιγραφή της όλης διαδικασίας.

---

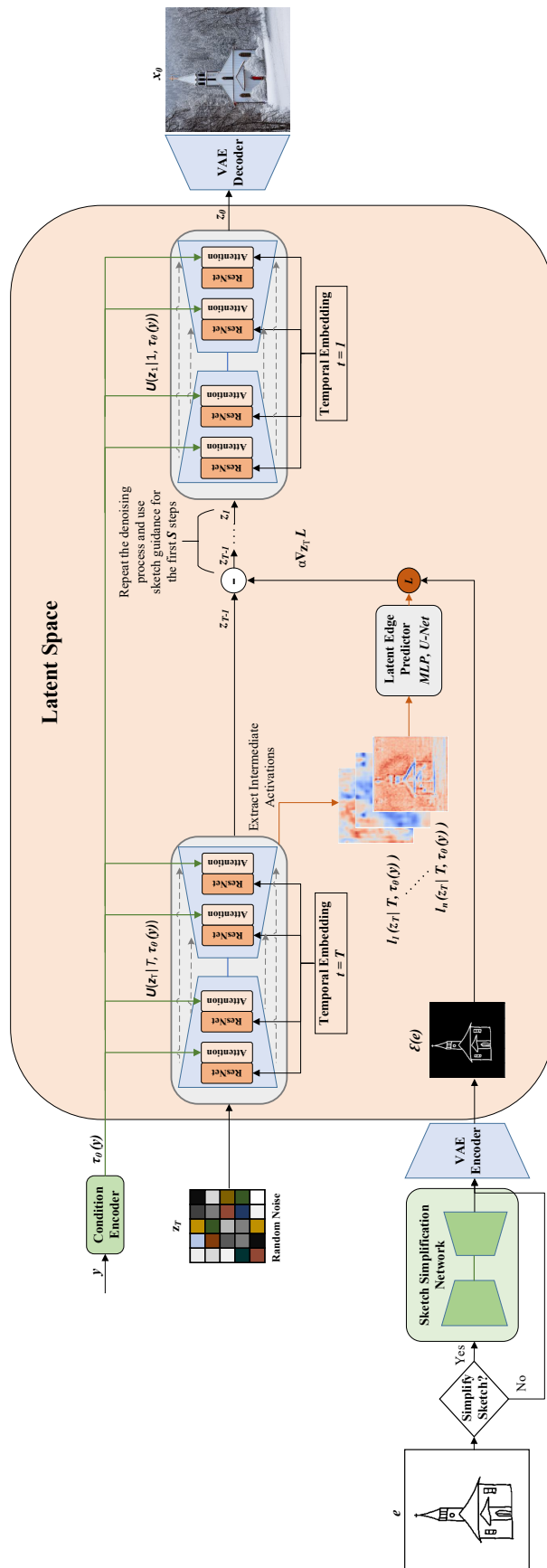
#### Αλγόριθμος 4.2: Sketch Guided Text-to-Image Synthesis

---

**Require:** scheduler ▷ scheduler to be used for the inverse diffusion  
**Require:** num\_inference\_steps (=  $T$ ) ▷ number of inference steps  
**Require:**  $\beta, S$  ▷ guidance strength, number of sketch guidance steps  
**Require:** ( $y, e$ ), sketch\_simplification ▷ prompt and target sketch  
**Require:** lep\_architecture ▷  $P$  (MLP) or  $U_{LEP}$  (UNET)

- 1:  $LEP = lep\_architecture$
- 2: **if** sketch\_simplification **then**
- 3:      $e = sketch\_simplification\_network(e)$
- 4: **end if**
- 5: Pass  $e$  through variational encoder to get  $\mathcal{E}(e)$
- 6: scheduler.set\_timesteps(num\_inference\_timesteps) ▷ set inference steps
- 7:  $z_T = random\_gaussian\_noise * scheduler.init\_noise\_sigma$  ▷ generate noise
- 8: **for**  $t$  in scheduler.timesteps **do**
- 9:     Pass  $z_T$  through the conditional U-Net to get  $U(z_T|t, \tau_\theta(y))$  ▷ noise estimation
- 10:     Calculate  $z_{t-1}$ :  $z_{t-1} = scheduler.prev\_sample(U(z_T|t, \tau_\theta(y)), t, z_t)$
- 11:     **if**  $T - t \leq S$  **then**
- 12:         Extract intermediate activations  $F(z_t|t, \tau_\theta(y))$
- 13:         Calculate  $\mathcal{L}(LEP(F(z_t|t, \tau_\theta(y))), \mathcal{E}(e))$  ▷ Eq.4.5
- 14:         Calculate  $\alpha$  ▷ Eq.4.7
- 15:         Update prediction  $z_{t-1}$ :  $z_{t-1} \leftarrow z_{t-1} - \alpha \nabla_{z_t} \mathcal{L}(LEP(F(z_t|t, \tau_\theta(y))), \mathcal{E}(e))$
- 16:     **end if**
- 17: **end for**
- 18: Pass final latent representation  $z_0$  through variational decoder to get  $x_0 = \mathcal{D}(z_0)$
- 19: **return**  $x_0$

---



Σχήμα 4.5: Σχηματική αναπαράσταση συνολικής προτεινόμενης μεθόδου για την καθοδηγούμενη από σκίτσα και κειμενικές περιγραφές σύνθεση εικόνων.

Μέρος 

Πειραματικό Μέρος

---

## Εισαγωγή

---

Στο πρώτο μέρος της παρούσα εργασίας αναλύθηκε το θεωρητικό υπόβαθρο και οι θεμελιώδεις αρχές λειτουργίας των μοντέλων διάχυσης, ενώ παρουσιάστηκε διεξοδικά η μεθοδολογία της καθοδηγούμενης από σκίτσα σύνθεσης εικόνων. Η ανάλυση αυτή ωστόσο, όντας θεωρητική, δεν παρείχε πληροφορίες σχετικά με την αριθμητική τιμή των διαφόρων παραμέτρων που υπεισέρχονται στη διαδικασία της καθοδηγούμενης σύνθεσης εικόνων και καθορίζουν το τελικό αποτέλεσμα.

Στο πλαίσιο αυτό, το παρόν μέρος της εργασίας είναι αφιερωμένο στις πρακτικές λεπτομέρειες της υλοποίησης της προτεινόμενης μεθόδου. Προς την κατεύθυνση αυτή, στο Κεφάλαιο 5, περιγράφεται ο τρόπος με τον οποίο κατασκευάζεται το σύνολο των δεδομένων εκπαίδευσης, το οποίο και χρησιμοποιείται για την εκπαίδευση των μοντέλων των latent edge predictors. Έπειτα, στα Κεφάλαια 6 και 7 παρουσιάζονται αποτελέσματα καθοδηγούμενης από σκίτσα σύνθεσης εικόνων με χρήση των δικτύων per-pixel MLP και U-Net ως latent edge predictors, αντίστοιχα. Ακολούθως, στο Κεφάλαιο 8 γίνεται μία ποιοτική και ποσοτική συγκριτική αξιολόγηση της επίδοσης των δύο αυτών μοντέλων, ως προς την ικανότητά τους να παράγουν ρεαλιστικές εικόνες, οι οποίες και ακολουθούν τα χωρικά περιγράμματα των σκίτσων αναφοράς. Τέλος, στο Κεφάλαιο 9 γίνεται μια συνοπτική ανακεφαλαίωση των όσων αναλύθηκαν και αναπτύχθηκαν στα προηγούμενα κεφάλαια και συνοψίζεται η συμπερασματολογία που προκύπτει από τα παρατιθέμενα αποτελέσματα.



## Κεφάλαιο 5

### Κατασκευή Συνόλου Δεδομένων Εκπαίδευσης

---

Απαραίτητη προϋπόθεση για την αποτελεσματική και επιτυχημένη εκπαίδευση του latent edge predictor, είτε αυτός βασίζεται στη χρήση του δικτύου MLP, είτε του δικτύου U-Net, είναι η ύπαρξη ενός πλήρους και αρκούντως μεγάλου συνόλου δεδομένων, το οποίο, βάσει της περιγραφής της διαδικασίας εκπαίδευσης (ενότητα 4.1), θα πρέπει να αποτελείται από τριπλέτες της μορφής  $(x, e, y)$ , όπου  $x$  μία εικόνα,  $y$  η κειμενική περιγραφή και  $e$  ο αντίστοιχος χάρτης ακμών, ήτοι το σκίτσο αναφοράς.

Προκειμένου λοιπόν να παραχθούν τριπλέτες της μορφής αυτής, αξιοποιείται το σύνολο εικόνων *ImageNet* [54]. Το dataset αυτό αποτελείται από μία πολύ μεγάλη συλλογή εικόνων, οι οποίες συνοδεύονται από τις αντίστοιχες επισημειώσεις των κλάσεων στις οποίες και ανήκουν. Από το dataset αυτό επιλέγονται συνολικά 40 εικόνες από 50 διαφορετικές κλάσεις και ως εκ τούτου προκύπτει ένα σύνολο **6000 εικόνων**, το οποίο και θα αποτελέσει το σύνολο των δεδομένων εκπαίδευσης.

Οι εικόνες αυτές ωστόσο, δε συνοδεύονται από τους αντίστοιχους χάρτες ακμών, οι οποίοι και είναι απαραίτητοι κατά τη διαδικασία της εκπαίδευσης. Για να εξαχθούν λοιπόν οι χάρτες αυτοί, χρησιμοποιείται το δίκτυο *PiDiNet* (Pixel Difference Networks for Efficient Edge Detection) [55], το οποίο και πρόκειται για ένα συνελκτικό δίκτυο εξαγωγής ακμών, βασιζόμενο στη λογική των pixel difference συνελίξεων. Μέσω της χρήσης του δικτύου αυτού, παράγονται συνολικά 6000 χάρτες ακμών, οι οποίοι και θα αποτελέσουν ουσιαστικά τα σκίτσα αναφοράς των αντίστοιχων εικόνων, από τις οποίες και προήλθαν. Στο σημείο αυτό αξίζει να επισημανθεί, ότι το δίκτυο PiDiNet παράγει χάρτες ακμών διαφορετικών εντάσεων, τούτέστιν δεν παράγει δυαδικές εικόνες, αλλά grayscale εικόνες, οι οποίες καταλαμβάνουν όλο το εύρος των χρωματικών εντάσεων. Προκειμένου λοιπόν να αποκτήσουμε δυαδικούς χάρτες ακμών, η έξοδος του PiDiNet τροποποιείται βάσει ενός καταφλίου τιμής 0.5, έτσι ώστε pixels με τιμή μικρότερη του 0.5 να μηδενίζονται, ενώ pixels με τιμή μεγαλύτερη ή ίση του 0.5 να λαμβάνουν τιμή ίση με 1. Τέλος, η κλάση στην οποία ανήκει η κάθε εικόνα, λειτουργεί ως κειμενική περιγραφή για τη σύνθεση της εικόνας αυτής, έτσι ώστε να ολοκληρωθεί ο σχηματισμός της ζητούμενης τριπλέτας. Συνολικά, η διαδικασία κατασκευής του συνόλου των δεδομένων εκπαίδευσης περιγράφεται στον Αλγόριθμο 5.1, ενώ στο Σχ.5.1 παρουσιάζονται μερικά δείγματα τριπλετών της μορφής  $(x, e, y)$ .

---

**Αλγόριθμος 5.1:** Dataset Creation


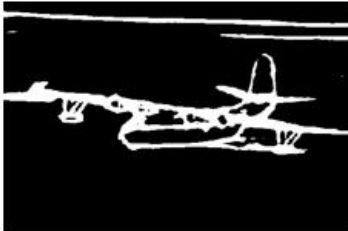




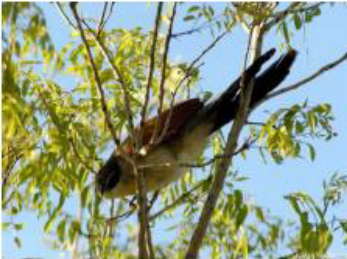

---

**Require:**  $x_i, i = 1, \dots, 6000$   $\triangleright$  6000 images from ImageNet

1: dataset\_triplets = [ ]

2: **for**  $i = 1, \dots, 6000$  **do**3:     Pass image  $x_i$  through PiDiNet and threshold to get  $e_i$ 4:     Set  $y_i =$  class name      $\triangleright$  Use class name as prompt5:     dataset\_triplets.append( $(x_i, e_i, y_i)$ )6: **end for**

---

| Image (x)   | Edge Map (e)   | Prompt (y)         |
|---|--|--------------------|
|   |   | <i>Flying boat</i> |
|  |  | <i>Castle</i>      |
|  |  | <i>Bed</i>         |
|  |  | <i>Coucal</i>      |

**Σχήμα 5.1:** Δείγματα τριπλετών  $(x, e, y)$  του συνόλου των δεδομένων εκπαίδευσης.

# Κεφάλαιο 6

## Αποτελέσματα - per-pixel MLP

---

Στο κεφάλαιο αυτό θα παρουσιαστούν αποτελέσματα καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων, όταν χρησιμοποιείται το δίκτυο του per pixel MLP ως latent edge predictor (ενότητα 4.1). Πριν ωστόσο γίνει η παράθεση των αποτελεσμάτων αυτών, θα πρέπει να προσδιοριστούν οι παράμετροι, οι οποίοι και χρησιμοποιήθηκαν για την εκπαίδευση του εν λόγω δικτύου.

### 6.1 Παράμετροι Εκπαίδευσης

Αρχικά λοιπόν και όσον αφορά στις γενικές παραμέτρους της εκπαίδευσης, το δίκτυο εκπαιδεύεται για **10 εποχές**, βάσει των συνολικά **6000 τριπλετών** της μορφής  $(x_i, e_i, y_i)$ , οι οποίες και προέκυψαν σύμφωνα με τη διαδικασία που περιγράφηκε στην ενότητα 5. Ως βελτιστοποιητής χρησιμοποιείται ο **Adam**, με ρυθμό μάθησης ίσο με **learning rate = 0.001** και χρήση scheduler **constant with warmup**, ενώ το **batch size** τίθεται ίσο με **16**.

Τέλος, ιδιαίτερως σημαντική είναι και η επιλογή του *Scheduler*. Οι schedulers καθορίζουν τη μεθοδολογία που χρησιμοποιείται τόσο για την επαναλαμβανόμενη προσθήκη θορύβου στα αρχικά δείγματα (εξίσωση 2.4), όσο και για τον υπολογισμό του δείγματος του προηγούμενου επιπέδου θορύβου κατά την αντίστροφη διαδικασία της αποθορυβοποίησης, βάσει της εξόδου του δικτύου εκτίμησης του θορύβου. Στα πλαίσια της παρούσας εργασίας, τόσο για την εκπαίδευση του per-pixel MLP latent edge predictor, όσο και για τη διαδικασία της αντίστροφης διάχυσης, χρησιμοποιείται ο **DDIMScheduler** της πλατφόρμας *HuggingFace*, ο οποίος υλοποιεί τη μεθοδολογία των DDIMs (ενότητα 2.4), η οποία επιταχύνει σημαντικά τη διαδικασία της αντίστροφης διάχυσης, μειώνοντας τον αριθμό των απαιτούμενων βημάτων που απαιτείται για την παραγωγή ρεαλιστικών δειγμάτων. Ως αρχική τιμή της παραμέτρου  $\beta$  κατά τη διαδικασία της αποθορυβοποίησης ορίζεται η  $\beta_T = 0.00085$  και ως τελική η  $\beta_1 = 0.012$ , ενώ χρησιμοποιείται linear variance schedule. Οι παράμετροι εκπαίδευσης του per-pixel MLP latent edge predictor παρουσιάζονται συγκεντρωτικά στον Πίνακα 6.1. Όσον αφορά στο χρόνο εκπαίδευσης, αυτός ανέρχεται περίπου στις 5 ώρες, ενώ χαρακτηριστικό είναι το ότι έπειτα από την όγδοη περίπου εποχή, δεν παρατηρείται ιδιαίτερη μεταβολή στη συνάρτηση απώλειας και ως εκ τούτου στις τιμές των βαρών του per-pixel MLP.

**Πίνακας 6.1:** Παράμετροι εκπαίδευσης per-pixel MLP latent edge predictor.

| Παράμετρος    | Τιμές  |
|---------------|--|
| epochs        | 10   |
| optimizer     | Adam   |
| learning rate | 0.001 με constant warmup   |
| batch size    | 16   |
| Scheduler     | DDIM $\left\{ \begin{array}{l} \beta_T = 0.00085 \\ \beta_1 = 0.012 \\ \text{linear variance schedule} \end{array} \right\}$ |

## 6.2 Παραδείγματα Σύνθεσης

Στο σημείο αυτό και έχοντας παρουσιάσει τη διαδικασία εκπαίδευσης του per-pixel MLP latent edge predictor, θα γίνει παράθεση ορισμένων ενδεικτικών αποτελεσμάτων της διαδικασίας της καθοδηγούμενης από σκίτα σύνθεσης εικόνων, με χρήση του εν λόγω δικτύου. Όπως έχει ήδη αναφερθεί, η όλη διαδικασία της σύνθεσης προϋποθέτει την εκ των προτέρων επιλογή των τιμών ορισμένων υπερπαραμέτρων, όπως φαίνεται και στον Αλγόριθμο 4.2. Στον Πίνακα 6.2 συνοψίζονται οι τιμές των παραμέτρων αυτών. Για την παραγωγή υψηλής ποιότητας εικόνων βάσει κειμενικών περιγραφών χρησιμοποιείται το μοντέλο [Stable Diffusion v1.5](#).

Από τον πίνακα αυτό, έχει παραληφθεί μία πολύ σημαντική παράμετρος, η οποία αφορά στην επιλογή των ενδιάμεσων στρωμάτων του δικτύου αποθορυβοποίησης U-Net του Stable Diffusion, από τα οποία εξάγονται οι ενεργοποιήσεις  $\{I_1 \dots I_n\}$ , βάσει των οποίων κατασκευάζεται το διάνυσμα  $F(z_t|t, \tau_\theta(y))$ , σύμφωνα με την εξίσωση 4.3. Τα layers αυτά, όπως έχει περιγραφεί και στην ενότητα 3.2.2, αποτελούνται από δύο επιμέρους layers, ένα ResNet layer και ένα cross-attention layer. Το πρώτο εξ αυτών αναλαμβάνει την επεξεργασία και την αξιοποίηση της πληροφορίας σχετικά με το εκάστοτε τρέχον χρονικό βήμα της αποθορυβοποίησης, ενώ το δεύτερο συσχετίζει τα tokens της κειμενικής περιγραφής εισόδου, με τη χωρική διάταξη της παραγόμενης εικόνας. Στα πλαίσια της υλοποίησης της υφιστάμενης εργασίας, όλες οι ενεργοποιήσεις λαμβάνονται από τις εξόδους των cross-attention layers του δικτύου U-Net και συγκεκριμένα:

- από το **contracting block** λαμβάνονται οι ενεργοποιήσεις από τις εξόδους των στρωμάτων **2, 4, 8**,
- από το **bottleneck** λαμβάνονται οι ενεργοποιήσεις από τις εξόδους των στρωμάτων **0, 1, 2** και
- από το **expansive block** λαμβάνονται οι ενεργοποιήσεις από τις εξόδους των στρωμάτων **2, 4, 8**.

**Πίνακας 6.2:** Παράμετροι διαδικασίας καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων, με χρήση του per-pixel MLP latent edge predictor.

| Παράμετρος                  | Τιμή   |
|-----------------------------|--|
| $\beta$                     | 1.6  |
| num_inference_steps ( $T$ ) | 250  |
| $S$                         | $0.5T = 125$   |
| Scheduler                   | DDIM $\left\{ \begin{array}{l} \beta_T = 0.00085 \\ \beta_1 = 0.012 \\ \text{linear variance schedule} \end{array} \right\}$ |

Συνολικά, λαμβάνονται **9 διανύσματα ενεργοποιήσεων**, οπότε και το διάνυσμα  $F$  της εξίσωσης 4.3 ισούται με,

$$F(z_t|t, \tau_\theta(\mathbf{y})) = [l_1(\mathbf{w}|t, \tau_\theta(\mathbf{y})), \dots, l_9(\mathbf{w}|t, \tau_\theta(\mathbf{y})), t, \sin(2\pi t), \dots, \sin(2\pi t \cdot 2^{-9})].$$

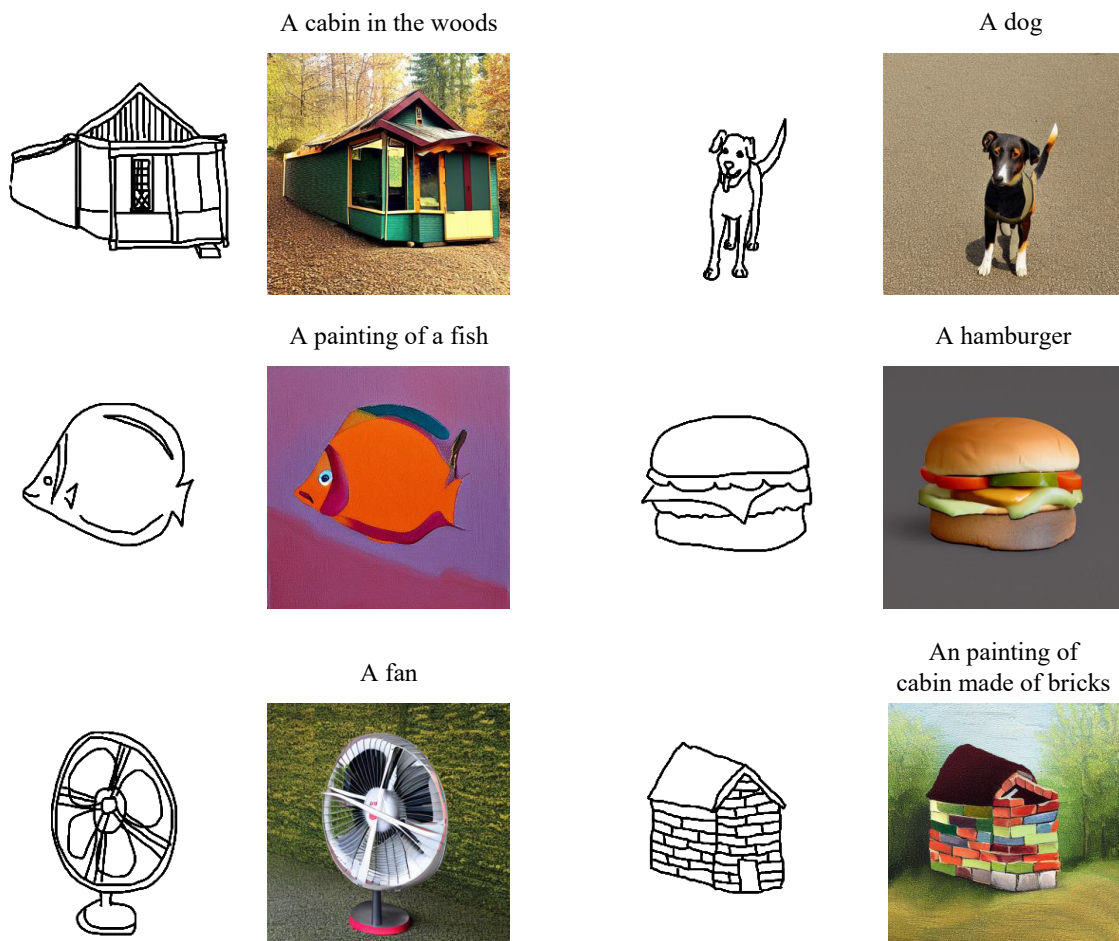
Το διάνυσμα αυτό είναι διαστάσεων  $[64, 64, c_F = 9320]$  και εμπεριέχει όλη τη χωρική και χρονική πληροφορία για την παραγόμενη εικόνα, κατά το τρέχον χρονικό βήμα  $t$  της διαδικασίας αποθορυβοποίησης.

Στα Σχ.6.1, 6.2 και 6.3 παρουσιάζονται ορισμένα παραδείγματα σύνθεσης εικόνων, με χρήση του per-pixel MLP latent edge predictor, βάσει των προαναφερθέντων τιμών των παραμέτρων. Προφανώς τα παραδείγματα αυτά είναι ενδεικτικά και αποσκοπούν στη γενική αποτύπωση της ικανότητας του δικτύου MLP να καθοδηγεί τη διαδικασία της σύνθεσης εικόνων, κάτω υπό διαφορετικές συνθήκες και κειμενικές περιγραφές. Τα σκίτσα αναφοράς προέρχονται από το σύνολο δεδομένων *Sketchy* [56].

Από τα παραδείγματα των Σχ.6.1, 6.2 και 6.3 εύκολα διαπιστώνει κανείς ότι η χρήση του per-pixel MLP latent edge predictor οδηγεί σε αρκετά ικανοποιητικά αποτελέσματα, με τις παραγόμενες εικόνες να ακολουθούν σε σημαντικό βαθμό το γεωμετρικό περίγραμμα των αντίστοιχων σκίτσων αναφοράς, διατηρώντας συγχρόνως τη φυσικότητα και το ρεαλισμό τους.

Μεγαλύτερη πιστότητα ως προς το περίγραμμα των σκίτσων αναφοράς, παρατηρείται στις περιπτώσεις όπου αυτά σχηματίζονται από καλά ορισμένες και ευδιάκριτες γραμμές, χωρίς έντονες επικαλύψεις και αλληλοδιασταυρώσεις. Γενικά, όσο πιο περίπλοκοι είναι οι σχηματισμοί και το πλήθος των γραμμών αυτών, οι παραγόμενες εικόνες είτε αδυνατούν να αποτυπώσουν με μεγάλη ακρίβεια τη γεωμετρική αυτή πολυπλοκότητα και λεπτομέρεια, όπως φαίνεται στην περίπτωση των εικόνων του κάστρου της δεύτερης σειράς του Σχ.6.2 και της εκκλησίας, της τρίτης σειράς του Σχ.6.3, είτε επιτυγχάνουν να αποτυπώσουν τη λεπτομέρεια αυτή σε βάρος του ρεαλισμού, όπως φαίνεται στην εικόνα του

ηλιοτροπίου στην τρίτη σειρά του Σχ.6.3. Ένα ακόμη σημαντικό συμπέρασμα το οποίο εξάγεται από τα παραδείγματα αυτά, είναι ότι σε περιπτώσεις όπου το σκίτσο αναφοράς παρουσιάζει σχετικά στατική γεωμετρία, υπό την έννοια των επαναλαμβανόμενων απλών γεωμετρικών σχημάτων, όπως για παράδειγμα ευθειών, οι παραγόμενες εικόνες τείνουν να ακολουθούν πιστά τη γεωμετρία αυτή, χάνοντας όμως και πάλι το στοιχείο του ρεαλισμού, κάτι το οποίο και φαίνεται στην περίπτωση της κλεψύδρας της τελευταίας γραμμής του Σχ.6.3. Τέλος, ένα ακόμη συμπέρασμα, το οποίο προκύπτει δευτερευόντως από τα παρατιθέμενα παραδείγματα και πρωτίστως από τα πειράματα που πραγματοποιήθηκαν, είναι ότι η συνολική διαδικασία της καθοδηγούμενης από σκίτσα σύνθεσης εικόνων είναι εν γένει πιο αποτελεσματική, σε περιπτώσεις όπου τα σκίτσα αναφοράς απεικονίζουν αντικείμενα και κατασκευές μεγαλύτερων διαστάσεων, όπως για παράδειγμα σπιτιών, κάστρων, πλοίων κ.ά. Αντιθέτως, σε περιπτώσεις όπου τα σκίτσα αναφέρονται σε μικρότερα αντικείμενα υψηλής λεπτομέρειας ή ζωντανούς οργανισμούς, τα αποτελέσματα δεν είναι το ίδιο ικανοποιητικά, γεγονός το οποίο οφείλεται στο υψηλό επίπεδο λεπτομέρειας και ακρίβειας που παρουσιάζουν οι απεικονίσεις αυτές στον πραγματικό κόσμο, έναντι του σχετικά απλοϊκού τρόπου με τον οποίο αποδίδονται μέσω των αντίστοιχων σκίτσων.



**Σχήμα 6.1:** Παραδείγματα καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων, με χρήση του δικτύου per-pixel MLP ως latent edge predictor. Για όλα τα παραδείγματα χρησιμοποιούνται οι παράμετροι του Πίνακα 6.2 κατά τη διαδικασία της αντίστροφης διάχυσης.

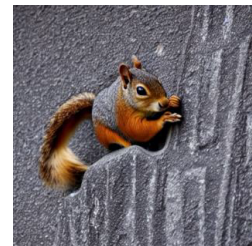


**Σχήμα 6.2:** Παραδείγματα καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων, με χρήση του δικτύου per-pixel MLP ως latent edge predictor. (συνέχεια)

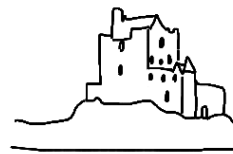
A castle at the edge of a hill



A squirrel



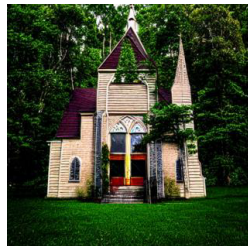
A camel



A castle on a hill next to a river



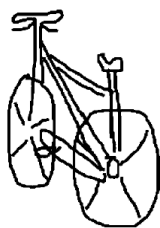
A church in the forest



A sunflower



A photo of an origami bicycle



A photo of windmill



An hourglass



An armor



Σχήμα 6.3: Παραδείγματα καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων, με χρήση του δικτύου per-pixel MLP ως latent edge predictor. (συνέχεια)



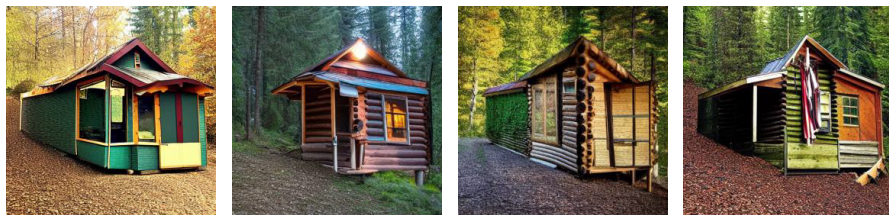
Πέραν των συμπερασμάτων που διεξήχθησαν έως τώρα, τα οποία και προκύπτουν άμεσα μέσω παρατήρησης των αποτελεσμάτων των Σχ.6.1, 6.2 και 6.3, αξίζει να γίνουν και ορισμένες περαιτέρω παρατηρήσεις, οι οποίες δεν είναι τόσο προφανείς και οι οποίες προκύπτουν μέσω επαναλαμβανόμενων δοκιμών και πειραματισμών.

### 6.2.1 Επίδραση Αρχικού Θορύβου

Αρχικά, η πιο σημαντική ίσως διαπίστωση για την όλη διαδικασία της καθοδηγούμενης από σκίτσα σύνθεσης εικόνων, είναι ότι αυτή επηρεάζεται σε πάρα πολύ μεγάλο βαθμό από την αρχικοποίηση του θορύβου. Όπως αναφέρθηκε και στην ενότητα 4.2, η όλη διαδικασία της καθοδηγούμενης αντίστροφης διάχυσης ξεκινά με τη δειγματοληψία τυχαίου Γκαουσιανού θορύβου, ο οποίος αποθορυβοποιείται διαδοχικά, ώστε να οδηγήσει στο επιθυμητό αποτέλεσμα. Η αρχικοποίηση αυτή, είναι υψίστης σημασίας για την επιτυχή καθοδήγηση των περιγραμμάτων της παραγόμενης εικόνας.

Ο αρχικός Γκαουσιανός θόρυβος, παρόλης της άτακτης και τυχαίας φύσης του, παρουσιάζει μία ενδογενή διάταξη, η οποία προφανώς δεν είναι ορατή και εντοπίσιμη και η οποία είναι ικανή είτε να διευκολύνει, είτε να δυσκολέψει σημαντικά τη διαδικασία της καθοδήγησης. Σε περιπτώσεις λοιπόν όπου η ενδογενής αυτή διάταξη του αρχικού θορύβου ευθυ-

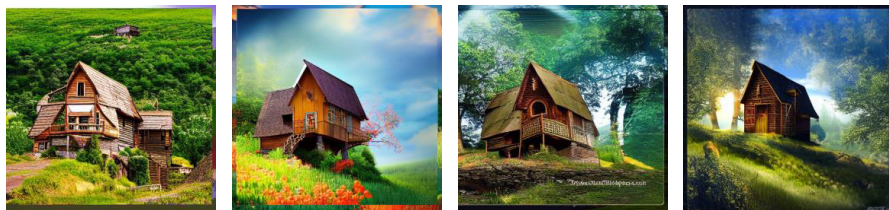
A cabin in the woods



A penguin



A fantasy picture of a wooden house on the hill in summer



**Σχήμα 6.4:** Παραδείγματα καθοδηγούμενης από σκίτσα σύνθεσης εικόνων, με χρήση του per-pixel MLP latent edge predictor και διαφορετικές αρχικοποιήσεις θορύβου.

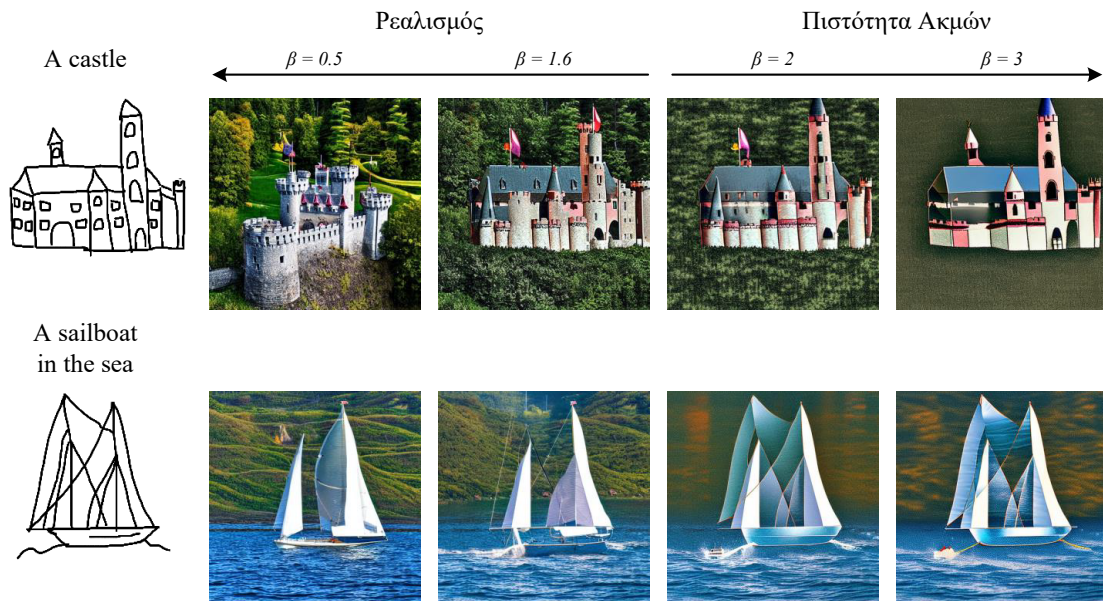
γραμμίζεται σε ένα βαθμό με την αντίστοιχη χωρική διάταξη των σκίτσων εισόδου, η διαδικασία της καθοδήγησης διευκολύνεται σημαντικά, με αποτέλεσμα να οδηγεί σε πολύ καλύτερα αποτελέσματα. Αντιθέτως, σε περιπτώσεις όπου οι διατάξεις αυτές απέχουν σημαντικά, η καθοδήγηση οδηγεί είτε σε εικόνες που δεν ακολουθούν τα αντίστοιχα σκίτσα, είτε σε εικόνες που ακολουθούν πιστά τα σκίτσα, χωρίς ωστόσο να είναι ρεαλιστικές. Στο Σχ.6.4 παρουσιάζονται ορισμένα αποτελέσματα σύνθεσης, όταν χρησιμοποιούνται διαφορετικές αρχικοποιήσεις θορύβου. Από τα αποτελέσματα αυτά εύκολα διαπιστώνεται, ότι ο αρχικός θόρυβος είναι καθοριστικής σημασίας για την πιστότητα και το ρεαλισμό των παραγόμενων εικόνων. Σε ορισμένες περιπτώσεις, οι διαφορετικές αρχικοποιήσεις οδηγούν σε ελαφρώς τροποποιημένα αποτελέσματα (εικόνες δεύτερης σειράς), ενώ σε άλλες περιπτώσεις οδηγούν σε εντελώς διαφορετικά αποτελέσματα, εκ των οποίων κάποια παρουσιάζουν ξεκάθαρη υπεροχή ως προς το ρεαλισμό και τη γεωμετρική απόδοση των περιγραμμάτων των σκίτσων αναφοράς (εικόνες πρώτης και τρίτης σειράς).

### 6.2.2 Επίδραση Sketch-guidance Strength ( $\beta$ )

Μία ακόμη πολύ σημαντική παρατήρηση αφορά στην επίδραση της παραμέτρου  $\beta$  στην όλη διαδικασία. Η παράμετρος αυτή, όπως έχει ήδη αναφερθεί, καθορίζει στο βαθμό στον οποίο η καθοδήγηση λόγω σκίτσου επηρεάζει τη σύνθεση των εικόνων, βάσει της εξίσωσης 4.7. Για όλα τα έως τώρα παραδείγματα, η τιμή της παραμέτρου αυτής είναι ίση με  $\beta = 1.6$ . Στο Σχ.6.5 παρουσιάζονται ορισμένα παραδείγματα σύνθεσης, με χρήση του per-pixel MLP latent edge predictor, για διαφορετικές τιμές της παραμέτρου  $\beta$ , διατηρώντας την ίδια κάθε φορά αρχικοποίηση θορύβου και τον ίδιο αριθμό βημάτων καθοδήγησης λόγω σκίτσου ( $S$ ). Από τα παραδείγματα αυτά διαπιστώνονται τα εξής:

- Αύξηση της τιμής της παραμέτρου  $\beta$  οδηγεί σε πιο πιστή αποτύπωση των περιγραμμάτων του σκίτσου αναφοράς, σε βάρος ωστόσο του ρεαλισμού των παραγόμενων εικόνων. Παρατηρείται δηλαδή μία αντιστρόφως ανάλογη σχέση μεταξύ ρεαλισμού και πιστότητας ακμών, γεγονός αναμενόμενο, αν αναλογιστεί κανείς τον τρόπο με τον οποίο η καθοδήγηση λόγω σκίτσου υπεισέρχεται στη διαδικασία της σύνθεσης και επηρεάζει τις παραγόμενες εικόνες.
- Ελάττωση της τιμής της παραμέτρου  $\beta$  οδηγεί σε εικόνες, οι οποίες παρουσιάζουν σημαντική απόκλιση από τα περιγράμματα των σκίτσων αναφοράς και οι οποίες διατηρούν ωστόσο τη φυσικότητα και το ρεαλισμό τους.

Συμπερασματικά, η επιλογή της παραμέτρου  $\beta$  είναι καθοριστική για την ποιότητα και την πιστότητα των παραγόμενων εικόνων και θα πρέπει με επιλέγεται με τρόπο, ώστε να εξασφαλίζεται μία σχετικά ισότιμη σχέση μεταξύ πιστότητας ακμών και ρεαλισμού. Τέλος, σημειώνεται ότι η παράμετρος αυτή παρουσιάζει έντονη εξάρτηση και από το πλήθος των βημάτων που χρησιμοποιούνται κατά τη διαδικασία της αντίστροφης διάχυσης. Πιο αναλυτικά, όσο μεγαλύτερος είναι ο αριθμός των βημάτων αποθορυβοποίησης, τόσο μικρότερη τιμή απαιτείται για την παράμετρο  $\beta$  και αυτό, διότι αυξάνεται αυτομάτως ο αριθμός των βημάτων στα οποία συντελείται η καθοδήγηση λόγω σκίτσου. Κατά αντίστοιχο τρόπο, όσο μικρότερος είναι ο αριθμός των βημάτων αποθορυβοποίησης, τόσο μικρότερος είναι και ο αριθμός στον οποίο συντελείται η καθοδήγηση λόγω σκίτσου. Επομένως, θα πρέπει να αυξηθεί η τιμή της παραμέτρου  $\beta$ , ώστε να καλυφθεί το κενό που προκαλείται στην καθοδήγηση λόγω του μειωμένου αριθμού βημάτων.

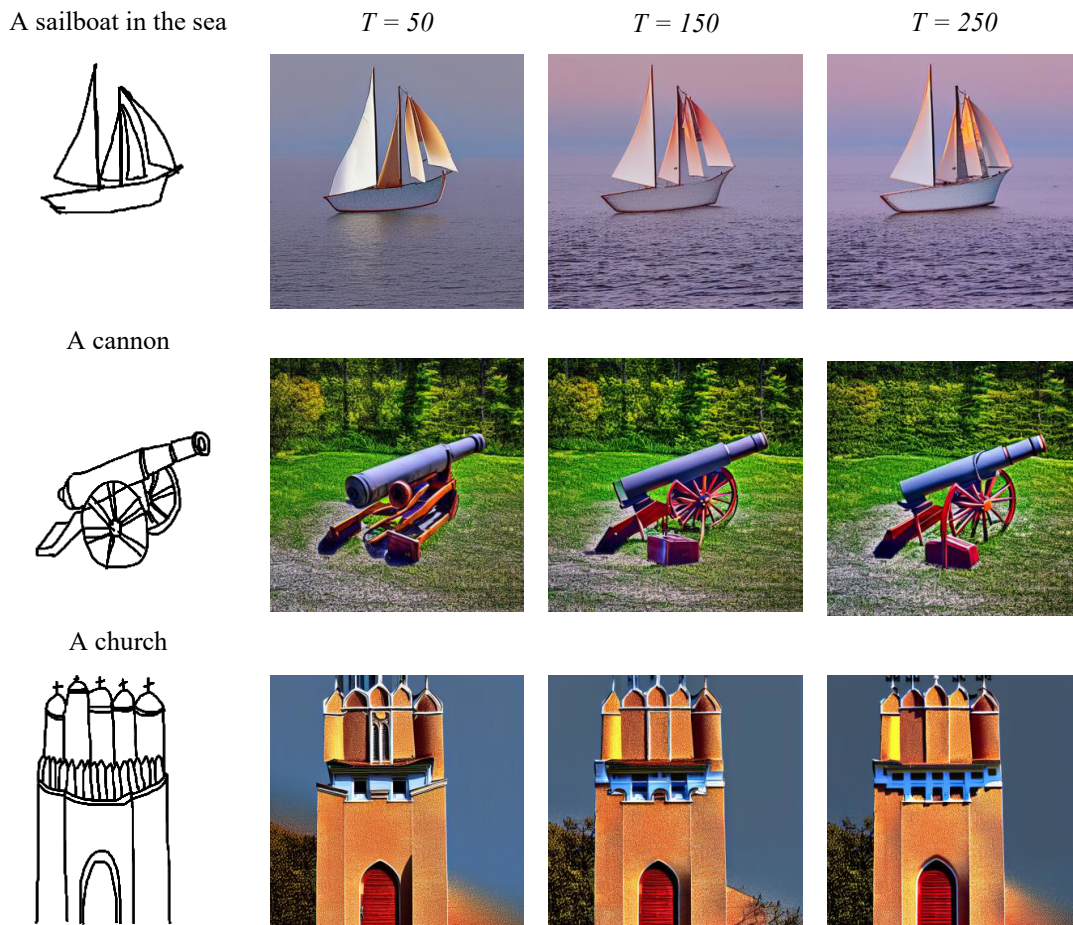


**Σχήμα 6.5:** Παραδείγματα καθοδηγούμενης από σκίτσα σύνθεσης εικόνων, με χρήση του per-pixel MLP latent edge predictor και διαφορετικές τιμές της παραμέτρου  $\beta$ .

### 6.2.3 Επίδραση Πλήθους Βημάτων Αντίστροφης Διαδικασίας Διάχυσης ( $T$ )

Τέλος, ειδική μνεία αξίζει να γίνει στον τρόπο με τον οποίο το πλήθος των βημάτων που χρησιμοποιείται κατά την αντίστροφη διαδικασία της διάχυσης επηρεάζει τα τελικά αποτελέσματα. Στη γενική περίπτωση των DDPMs, απαιτούνται συνολικά  $T$  βήματα κατά τη διαδικασία της αντίστροφης διάχυσης προκειμένου να παραχθούν νέες εικόνες, όπου  $T$  ο συνολικός αριθμός των βημάτων-επιπέδων θορύβου που χρησιμοποιήθηκαν κατά την προς τα εμπρός διαδικασία διάχυσης. Παρά ταύτα, καθώς ο αριθμός των βημάτων αυτών μπορεί να είναι απαγορευτικά μεγάλος, μπορεί να χρησιμοποιηθεί ένα υποσύνολο αυτού, αναλόγως πάντα και της μεθόδου που χρησιμοποιείται για την πρόβλεψη του δείγματος του προηγούμενου επιπέδου θορύβου. Στα πλαίσια της παρούσα εργασίας, χρησιμοποιείται, όπως σημειώθηκε και στην ενότητα 6.1, η μέθοδος των DDIMs για την υλοποίηση της αντίστροφης διαδικασίας διάχυσης, με αριθμό βημάτων ίσο με 250. Προκειμένου λοιπόν να διαπιστωθεί η επίδραση του αριθμού αυτού των βημάτων στην όλη διαδικασία, στο Σχ.6.6 παρουσιάζονται ορισμένα παραδείγματα σύνθεσης, τα οποία έχουν προκύψει με χρήση διαφορετικού αριθμού βημάτων αποθορυβοποίησης, διατηρώντας την ίδια κάθε φορά αρχικοποίηση θορύβου, τον ίδιο αριθμό βημάτων καθοδήγησης λόγω σκίτσου ( $S$ ) και την ίδια τιμή της παραμέτρου  $\beta$  ( $= 1.6$ ).

Από τα παραδείγματα του Σχ.6.6 συμπεραίνεται ότι, αύξηση του πλήθους των βημάτων της διαδικασίας αποθορυβοποίησης οδηγεί σε πιο ρεαλιστικά και πιστά ως προς τα σκίτσα αποτελέσματα, ενώ μείωση του αριθμού αυτού έχει ως αποτέλεσμα οι παραγόμενες εικόνες να αποκλίνουν ως προς τη γεωμετρία των σκίτσων και να υστερούν ως προς την ευκρίνεια και τον ρεαλισμό. Προφανώς, όσο μεγαλύτερος είναι ο αριθμός των βημάτων αποθορυβοποίησης, τόσο υψηλότερος είναι και ο συνολικός χρόνος που απαιτείται για



**Σχήμα 6.6:** Παραδείγματα καθοδηγούμενης από σκίτσα σύνθεσης εικόνων, με χρήση του per-pixel MLP latent edge predictor και διαφορετικές τιμές αριθμού βημάτων αποθορυβοποίησης.

την παραγωγή δειγμάτων. Ενδεικτικά αναφέρεται, ότι για την παραγωγή μίας εικόνας σε κάρτα γραφικών *NVIDIA T4*, απαιτούνται συνολικά περίπου 50 δευτερόλεπτα στην περίπτωση των 50 βημάτων και περίπου 250 δευτερόλεπτα στην περίπτωση των 250 βημάτων, δηλαδή περίπου πενταπλάσιος χρόνος. Ως εκ τούτου, παρατηρείται και πάλι ένα trade-off μεταξύ της ποιότητας των συνθετικών εικόνων και του αντίστοιχου χρόνου που απαιτείται για την παραγωγή τους. Ο χρόνος αυτός είναι άρρηκτα συνδεδεμένος με τις ικανότητες του hardware που υπάρχει διαθέσιμο κάθε φορά για την υλοποίηση της διαδικασίας της καθοδηγούμενης από σκίτσα σύνθεσης εικόνων. Κατά συνέπεια και λόγω της άμεσης αυτής εξάρτησης, ο αριθμός των βημάτων αποθορυβοποίησης θα πρέπει να επιλέγεται κάθε φορά βάσει του διαθέσιμου hardware, αλλά και της επιθυμητής ποιότητας και πιστότητας των παραγόμενων εικόνων.

### 6.3 Γενικά συμπεράσματα

Συγκεφαλαιώνοντας, στις προηγούμενες ενότητες παρατέθηκαν παραδείγματα καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων, με χρήση του δικτύου per-pixel MLP ως latent edge predictor και αναλύθηκαν οι διάφοροι παράμετροι που υπεισέρχονται στη διαδικασία της σύνθεσης, καθώς και ο τρόπος με τον οποίο την επηρεάζουν.

Βάσει των αποτελεσμάτων αυτών, διαπιστώνεται ότι η όλη διαδικασία είναι αρκετά αποτελεσματική και οι παραγόμενες εικόνες ακολουθούν σε ικανοποιητικό βαθμό τη γεωμετρία των αντίστοιχων σκίτσων αναφοράς, ενώ παράλληλα συμμορφώνονται με το νοηματικό περιεχόμενο των κειμενικών περιγραφών. Συγχρόνως, τα αποτελέσματα φανερώνουν την ευαισθησία της όλης διαδικασίας, στις τιμές των επιμέρους υπερπαραμέτρων. Ιδιαίτερως σημαντικές κρίνονται οι παράμετροι του αριθμού των βημάτων αποθορυβοποίησης και του βαθμού στον οποίο η καθοδήγηση λόγω σκίτσου επηρεάζει και μεταβάλλει τις προβλέψεις στα ενδιάμεσα στάδια της αποθορυβοποίησης (παράμετρος  $\beta$ ). Η επιλογή των τιμών των παραμέτρων αυτών πρέπει να γίνεται με τρόπο, ώστε οι συνθετικές εικόνες να διατηρούν το ρεαλισμό και την ευκρίνειά τους, αποτυπώνοντας συγχρόνως και κατά το δυνατόν τη γεωμετρία των σκίτσων αναφοράς. Αυτό το trade-off μεταξύ ρεαλισμού και πιστότητας στην αποτύπωση των ακμών των σκίτσων, καθιστά την όλη διαδικασία ιδιαίτερως επιρρεπή και στις αρχικοποιήσεις του τυχαίου θορύβου, ο οποίος, βάσει των ενδογενών χωρικών χαρακτηριστικών του, μπορεί να διευκολύνει ή να παρεμποδίσει σημαντικά την όλη διαδικασία της σύνθεσης. Τέλος, ζωτικής σημασίας είναι και ο παράγοντας του χρόνου, ο οποίος αυξάνεται αναλογικά με τον αριθμό των επιλεγμένων βημάτων αποθορυβοποίησης. Στην περίπτωση όπου ως latent edge predictor χρησιμοποιείται το per-pixel MLP, ο αριθμός των βημάτων αποθορυβοποίησης κατά τον οποίο τα παραγόμενα αποτελέσματα παρουσιάζουν μια σχετική σταθερότητα ως προς το ρεαλισμό και την πιστότητα των ακμών των σκίτσων ισούται με 150, με αύξηση του αριθμού αυτού να συνεπάγεται αυτομάτως περαιτέρω βελτίωση της απόδοσης της όλης διαδικασίας. Για τον αριθμό αυτό των βημάτων, απαιτούνται περίπου 3 λεπτά για την παραγωγή ενός δείγματος, σε κάρτα γραφικών *NVIDIA T4*, ενώ για μικρότερο αριθμό βημάτων, και κυρίως για αριθμό μικρότερο του 100 τα αποτελέσματα υστερούν σημαντικά σε ρεαλισμό και γεωμετρική πιστότητα. Αυτό έχει ως αποτέλεσμα, η όλη διαδικασία της σύνθεσης, αν και αποτελεσματική, να είναι δέσμια του σχετικά μεγάλου αριθμού βημάτων αποθορυβοποίησης και ως εκ τούτου ιδιαίτερως χρονοβόρα, ειδικά εάν αναλογιστεί κανείς ότι για την παραγωγή κάποιου πολύ καλού δείγματος μπορεί η διαδικασία να επαναληφθεί για διαφορετικές αρχικοποιήσεις θορύβου.

Στο επόμενο κεφάλαιο θα επαναληφθεί η διαδικασία του παρόντος κεφαλαίου, χρησιμοποιώντας ωστόσο το προτεινόμενο δίκτυο U-Net ως latent edge predictor και θα εξαχθούν αντίστοιχα συμπεράσματα για την αποτελεσματικότητά του, καθώς και για το κατά πόσο αυτό βελτιώνει την όλη διαδικασία της σύνθεσης σε σχέση με το per-pixel MLP.

# Κεφάλαιο 7

## Αποτελέσματα - U-Net

Στο παρόν κεφάλαιο θα παρουσιαστούν και πάλι αποτελέσματα καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων, χρησιμοποιώντας αυτή τη φορά την προτεινόμενη αρχιτεκτονική τύπου U-Net (ενότητα 4.3) ως latent edge predictor. Η διάρθρωση του κεφαλαίου είναι αντίστοιχη με αυτή του Κεφαλαίου 6, οπότε και θα παρουσιαστούν αρχικά ορισμένα αποτελέσματα και ακολούθως θα μελετηθεί ο τρόπος με τον οποίο οι διάφορες υπερπαραμέτροι επηρεάζουν την όλη διαδικασία.

Όσον αφορά στη διαδικασία εκπαίδευσης και προκειμένου να υπάρχει μία αμεροληψία ως προς τη μετέπειτα σύγκριση της απόδοσης των δύο αρχιτεκτονικών, το δίκτυο U-Net εκπαιδεύεται με τον ίδιο ακριβώς τρόπο που χρησιμοποιήθηκε για την εκπαίδευση του per-pixel MLP. Ως δεδομένα εκπαίδευσης χρησιμοποιούνται οι συνολικά 6000 τριπλέτες της μορφής  $(x_i, e_i, y_i)$ , οι οποίες και προέρχονται από το σύνολο δεδομένων ImageNet, ενώ το δίκτυο εκπαιδεύεται συνολικά για 10 εποχές. Για λόγους πληρότητας, στον Πίνακα 7.1 συνοψίζονται οι τιμές των παραμέτρων εκπαίδευσης του U-Net latent edge predictor.

**Πίνακας 7.1:** Παράμετροι εκπαίδευσης U-Net latent edge predictor.

| Παράμετρος    | Τιμές  |
|---------------|--|
| epochs        | 10   |
| optimizer     | Adam   |
| learning rate | 0.001 με constant warmup   |
| batch size    | 16   |
| Scheduler     | DDIM $\left\{ \begin{array}{l} \beta_T = 0.00085 \\ \beta_1 = 0.012 \\ \text{linear variance schedule} \end{array} \right\}$ |

## 7.1 Παραδείγματα Σύνθεσης

Στο σημείο αυτό, θα γίνει παράθεση ορισμένων ενδεικτικών αποτελεσμάτων της διαδικασίας της καθοδηγούμενης από σκίτα σύνθεσης εικόνων, με χρήση του δικτύου U-Net ως latent edge predictor. Όσον αφορά στις τιμές των παραμέτρων που χρησιμοποιούνται κατά τη διαδικασία της σύνθεσης, αυτές είναι ίδιες με αυτές της περίπτωσης του per-pixel MLP latent edge predictor, με μία ουσιώδη και βασική διαφορά. Όπως αναλύθηκε και στην ενότητα 4.3, ο λόγος για τον οποίο επιλέγεται η αντικατάσταση της βασικής αρχιτεκτονικής του per-pixel MLP με την αρχιτεκτονική τύπου U-Net, είναι η προσπάθεια εξαγωγής χωρικών σχέσεων των pixels των ενεργοποιήσεων των ενδιάμεσων στρωμάτων του Stable Diffusion και η μετέπειτα χρήση τους για την κατασκευή των προβλεπόμενων χαρτών ακμών.

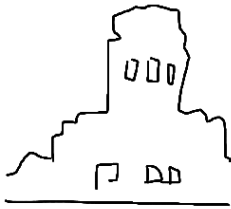
Στο πλαίσιο αυτό και έπειτα από αρκετούς πειραματισμούς διαπιστώθηκε, ότι η επιπλέον αυτή χωρική γνώση που προσφέρει η αρχιτεκτονική τύπου U-Net, οδηγεί στην παραγωγή πολύ καλών αποτελεσμάτων, τα οποία ακολουθούν ικανοποιητικά τα σκίτσα εισόδου και για τα οποία, ο αριθμός των **50 βημάτων** αποθορυβοποίησης εξασφαλίζει σε πολύ μεγάλο βαθμό την πιστότητα και το ρεαλισμό, κάτι το οποίο δεν παρατηρείται στην περίπτωση του per-pixel MLP. Επομένως, κρίνεται σκόπιμο να γίνει παράθεση των αποτελεσμάτων, με χρήση 50 βημάτων αποθορυβοποίησης, προκειμένου ακριβώς να αποδειχθεί η υπεροχή αυτή του δικτύου U-Net σε σχέση με το per-pixel MLP. Πιο διεξοδική και αναλυτική σύγκριση της επίδοσης των δύο δικτύων γίνεται σε επόμενο κεφάλαιο, ενώ στο παρόν, η προσοχή εστιάζεται κυρίως στην καθαυτή επίδοση του U-Net latent edge predictor.

Στον Πίνακα 7.2 συνοψίζονται οι τιμές των παραμέτρων της διαδικασίας της καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων, ενώ στα Σχ.7.1 και 7.2 παρατίθενται ορισμένα παραδείγματα σύνθεσης, με χρήση πάντοτε του U-Net latent edge predictor. Όπως και στην περίπτωση του MLP latent edge predictor, για την παραγωγή υψηλής ποιότητας εικόνων βάσει κειμενικών περιγραφών χρησιμοποιείται το μοντέλο [Stable Diffusion v1.5](#).

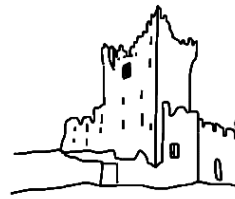
**Πίνακας 7.2:** Παράμετροι διαδικασίας καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεση εικόνων, με χρήση του U-Net latent edge predictor.

| Παράμετρος                  | Τιμή   |
|-----------------------------|--|
| $\beta$                     | 1.6  |
| num_inference_steps ( $T$ ) | 50   |
| $S$                         | $0.5T = 25$  |
| Scheduler                   | DDIM $\left\{ \begin{array}{l} \beta_T = 0.00085 \\ \beta_1 = 0.012 \\ \text{linear variance schedule} \end{array} \right\}$ |

A castle next to a river



A mosaic painting of a castle next to a river



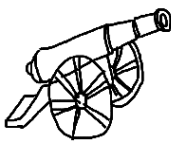
A sailboat in the sea



An hourglass



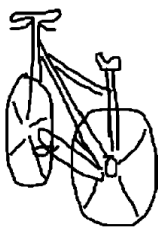
A cannon



A painting of a shark



A wooden bicycle



A table



A windmill

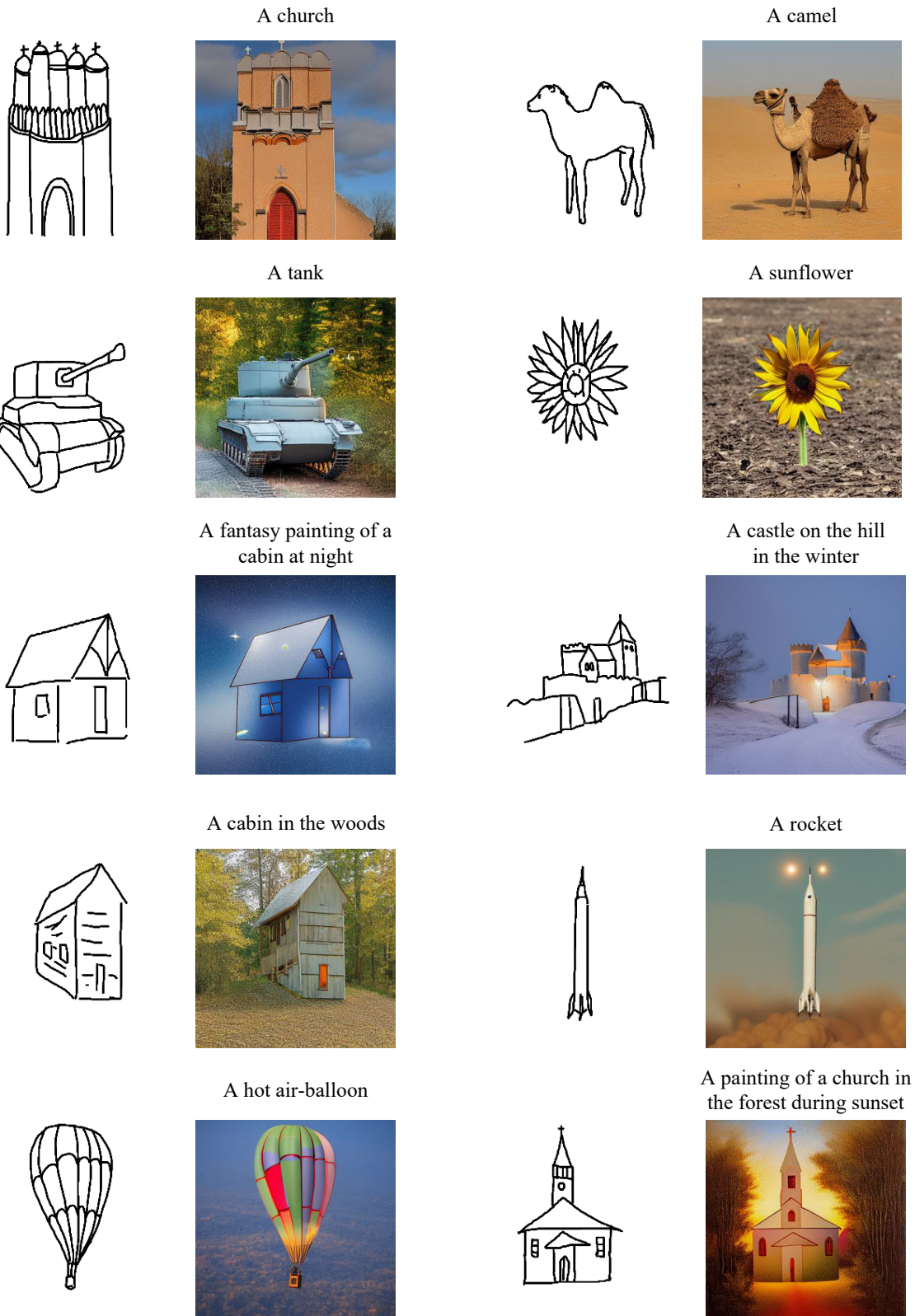


A church in the winter



**Σχήμα 7.1:** Παραδείγματα καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων, με χρήση του δικτύου U-Net ως latent edge predictor. Για όλα τα παραδείγματα χρησιμοποιούνται οι παράμετροι του Πίνακα 7.2 κατά τη διαδικασία της αντίστροφης διάχυσης.





**Σχήμα 7.2:** Παραδείγματα καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων, με χρήση του δικτύου U-Net ως latent edge predictor. (συνέχεια)

Από τα αποτελέσματα των Σχ.7.1 και 7.2 συμπεραίνεται, ότι η χρήση της αρχιτεκτονικής τύπου U-Net ως latent edge predictor οδηγεί σε πολύ καλά αποτελέσματα, με τις παραγόμενες εικόνες να συμμορφώνονται τόσο με τα περιγράμματα των σκίτσων εισόδου, όσο και με τις αντίστοιχες κειμενικές περιγραφές. Όπως και στην περίπτωση του per-pixel MLP, έτσι και τώρα, καλύτερη πιστότητα ως προς τα σκίτσα αναφοράς παρατηρείται σε περιπτώσεις όπου αυτά αποτελούνται από ευδιάκριτες και μη επικαλυπτόμενες γραμμές, όπως για παράδειγμα στην περίπτωση της κλεψύδρας και του τραπέζιού του Σχ.7.1, του πυραύλου και της εκκλησίας του Σχ.7.2. Ακόμα όμως και σε συνθήκες όπου τα σκίτσα αναφοράς χαρακτηρίζονται από σχετικά υψηλότερη πολυπλοκότητα ως προς τη χωρική διάταξη των γραμμών που τα απαρτίζουν, τα αποτελέσματα είναι αρκούντως ικανοποιητικά και αποδίδουν σε πολύ μεγάλο βαθμό τη χωρική αυτή πολυπλοκότητα. Στο πλαίσιο αυτό, χαρακτηριστικά παραδείγματα αποτελούν ο ανεμόμυλος και το ποδήλατο του Σχ.7.1, το ηλιοτρόπιο και το αερόστατο του Σχ.7.2, στα οποία οι συνθετικές εικόνες αποδίδουν πολύ ικανοποιητικά τις μικρές λεπτομέρειες των σκίτσων αναφοράς.

Πέραν των προαναφερθέντων, το πιο σημαντικό συμπέρασμα το οποίο εξάγεται από τα ανωτέρω αποτελέσματα, είναι ότι στην περίπτωση της χρήσης του U-Net ως latent edge predictor, ο αριθμός των 50 βημάτων αποθορυβοποίησης είναι αρκετός, ώστε να παραχθούν συνθετικές εικόνες υψηλού επιπέδου ρεαλισμού και φυσικότητας, με την κάθε εικόνα να απαιτεί περίπου 50 δευτερόλεπτα για την παραγωγή της, με χρήση μίας κάρτας γραφικών *NVIDIA T4*. Η παρατήρηση αυτή είναι ιδιαίτερης σημασίας, αφού ο χρόνος αποτελεί πάντα βασικό παράγοντα για την αξιολόγηση ενός συστήματος. Περισσότερα για τη χρονική αυτή βελτίωση της διαδικασίας σύνθεσης αναφέρονται στο Κεφάλαιο 8, όπου και γίνεται μία συγκριτική αξιολόγηση των δύο αρχιτεκτονικών του latent edge predictor.

Στις επόμενες ενότητες γίνεται μελέτη της επίδρασης των επιμέρους υπερπαραμέτρων και αρχικοποιήσεων στη διαδικασία της καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων.

### 7.1.1 Επίδραση Αρχικού Θορύβου

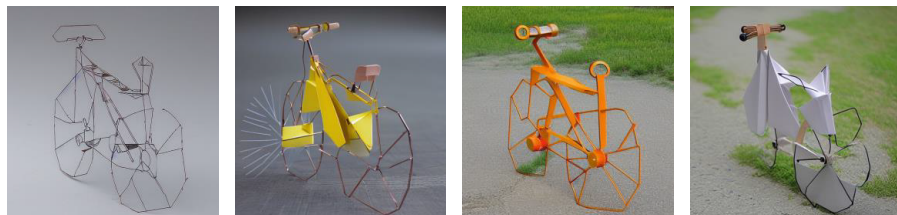
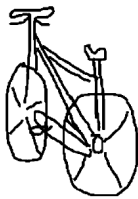
Όπως και στην περίπτωση του per-pixel MLP latent edge predictor, θα εξεταστεί αρχικά ο τρόπος με τον οποίο η αρχικοποίηση του Γκαουσιανού θορύβου επηρεάζει και μεταβάλλει τα τελικά αποτελέσματα της σύνθεσης. Για το σκοπό αυτό, στο Σχ.7.3 παρουσιάζονται ορισμένα αποτελέσματα σύνθεσης, όταν χρησιμοποιούνται διαφορετικές αρχικοποιήσεις θορύβου. Σε όλα τα αποτελέσματα, οι υπόλοιπες τιμές των υπερπαραμέτρων (Πίνακας 7.2) διατηρούνται σταθερές.

Από τα παραδείγματα αυτά, εύκολα αντιλαμβάνεται κανείς ότι και στην περίπτωση του U-Net latent edge predictor, η αρχικοποίηση του Γκαουσιανού θορύβου επηρεάζει σε πολύ μεγάλο βαθμό τα τελικά αποτελέσματα της σύνθεσης. Έτσι, σε περιπτώσεις όπου ο αρχικός θόρυβος παρουσιάζει ενδογενή χωρική διάταξη παρόμοια με αυτή του αντίστοιχου σκίτσου αναφοράς, η εικόνα που προκύπτει είναι πολύ πιο ρεαλιστική και πιστή ως προς τις ακμές του σκίτσου αυτού, ενώ σε διαφορετική περίπτωση, η διαδικασία της σύνθεσης δυσχεραίνεται σημαντικά, και η τελική εικόνα μπορεί να παρουσιάζει είτε απόκλι-

A cabin in the woods



An origami bicycle



A sailboat in the sea



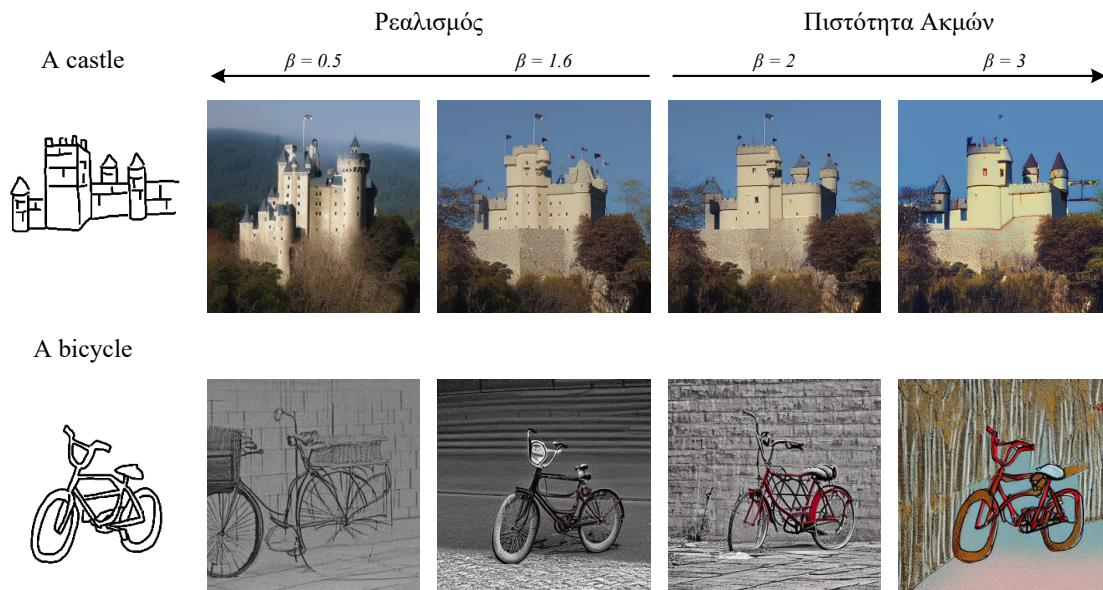
**Σχήμα 7.3:** Παραδείγματα καθοδηγούμενης από σκίτσα text-to-image σύνθεσης, με χρήση του U-Net latent edge predictor και διαφορετικές αρχικοποιήσεις θορύβου.

ση από το αρχικό σκίτσο (πρώτη συνθετική εικόνα πρώτης σειράς), είτε έλλειψη σε επίπεδο ρεαλισμού (τρίτη συνθετική εικόνα τρίτης σειράς). Σε κάθε περίπτωση, επιβεβαιώνεται και για τον U-Net latent edge predictor, ότι η αρχικοποίηση του Γκαουσιανού θορύβου αποτελεί ίσως το πιο μείζονος σημασίας στάδιο της όλης διαδικασίας. Ως εκ τούτου, σε ορισμένες περιπτώσεις και προκειμένου να ληφθούν ικανοποιητικά αποτελέσματα απαιτείται η επανάληψη της διαδικασίας της σύνθεσης με την ίδια κειμενική περιγραφή, για διαφορετικές αρχικοποιήσεις θορύβου.

### 7.1.2 Επίδραση Sketch-guidance Strength ( $\beta$ )

Επόμενη παράμετρος προς μελέτη, είναι αυτή του βαθμού επίδρασης της καθοδήγησης λόγω σκίτσου στη διαδικασία της σύνθεσης. Στο Σχ.7.4 παρουσιάζονται ορισμένα παραδείγματα σύνθεσης, με χρήση του U-Net latent edge predictor, για διαφορετικές τιμές της παραμέτρου  $\beta$ , διατηρώντας την ίδια κάθε φορά αρχικοποίηση θορύβου και τον ίδιο αριθμό βημάτων καθοδήγησης λόγω σκίτσου ( $S$ ). Σε κάθε περίπτωση, ο αριθμός των βημάτων αποθορυβοποίησης παραμένει σταθερός και ίσος με  $T = 50$ .

Όπως λοιπόν και στην περίπτωση του per-pixel MLP, καθώς η τιμή της παραμέτρου  $\beta$  αυξάνεται, οι παραγόμενες εικόνες ακολουθούν πιο πιστά το σκίτσο αναφοράς, σε βάρος ωστόσο του ρεαλισμού τους. Αντιθέτως, μείωση της τιμής της παραμέτρου  $\beta$  οδηγεί σε



**Σχήμα 7.4:** Παραδείγματα καθοδηγούμενης από σκίτσα σύνθεσης εικόνων, με χρήση του U-Net latent edge predictor και διαφορετικές τιμές της παραμέτρου  $\beta$ .

παραγόμενες εικόνες, οι οποίες αποκλίνουν σημαντικά από τη γεωμετρία των δοθέντων σκίτσων και οι οποίες δεν είναι υποχρεωτικά πιο ρεαλιστικές. Στο πλαίσιο αυτό, χαρακτηριστική είναι η περίπτωση των εικόνων της δεύτερης σειράς, όπου η πρώτη εξ αυτών, αν και λιγότερο δεσμευμένη ως προς την ακολουθία του αντίστοιχου σκίτσου, αποτυγχάνει να αποδώσει με ρεαλιστικό τρόπο την εικόνα ενός ποδηλάτου.

Συνολικά και στην περίπτωση του U-Net latent edge predictor, η τιμή της παραμέτρου  $\beta$  επηρεάζει σε πολύ μεγάλο βαθμό τα τελικά αποτελέσματα της διαδικασίας της σύνθεσης και θα πρέπει να επιλέγεται με τρόπο, ώστε να εξασφαλίζεται κατά το δυνατόν μία μορφή ισορροπίας μεταξύ ρεαλισμού και πιστότητας των ακμών του σκίτσου αναφοράς. Τέλος, σημειώνεται για ακόμη μία φορά, ότι η τιμή της παραμέτρου αυτής παρουσιάζει έντονη αρνητική συσχέτιση με τον αριθμό των βημάτων αποθορυβοποίησης, με αύξηση του αριθμού αυτού να συνεπάγεται και κάποια αντίστοιχη μείωση του βαθμού στον οποίο η καθοδήγηση λόγω σκίτσου επηρεάζει τη διαδικασία της σύνθεσης.

### 7.1.3 Επίδραση Πλήθους Βημάτων Αντίστροφης Διαδικασίας Διάχυσης ( $T$ )

Η μελέτη των επιμέρους υπερπαραμέτρων ολοκληρώνεται με την παράμετρο του πλήθους των βημάτων της αντίστροφης διαδικασίας διάχυσης. Στο Σχ.7.5 παρουσιάζονται ορισμένα παραδείγματα σύνθεσης, τα οποία έχουν προκύψει με χρήση διαφορετικού αριθμού βημάτων αποθορυβοποίησης, διατηρώντας την ίδια κάθε φορά αρχικοποίηση θορύβου, τον ίδιο αριθμό βημάτων καθοδήγησης λόγω σκίτσου ( $S$ ) και την ίδια τιμή της παραμέτρου  $\beta$  ( $= 1.6$ ).

Τα αποτελέσματα αυτά, φανερώνουν μία πολύ σημαντική ιδιότητα της αρχιτεκτονικής του U-Net latent edge predictor. Πιο αναλυτικά, βάσει και των τριών παραδειγμάτων,

A sailboat in the sea

 $T = 50$  $T = 150$  $T = 250$ 

A windmill



An old style table



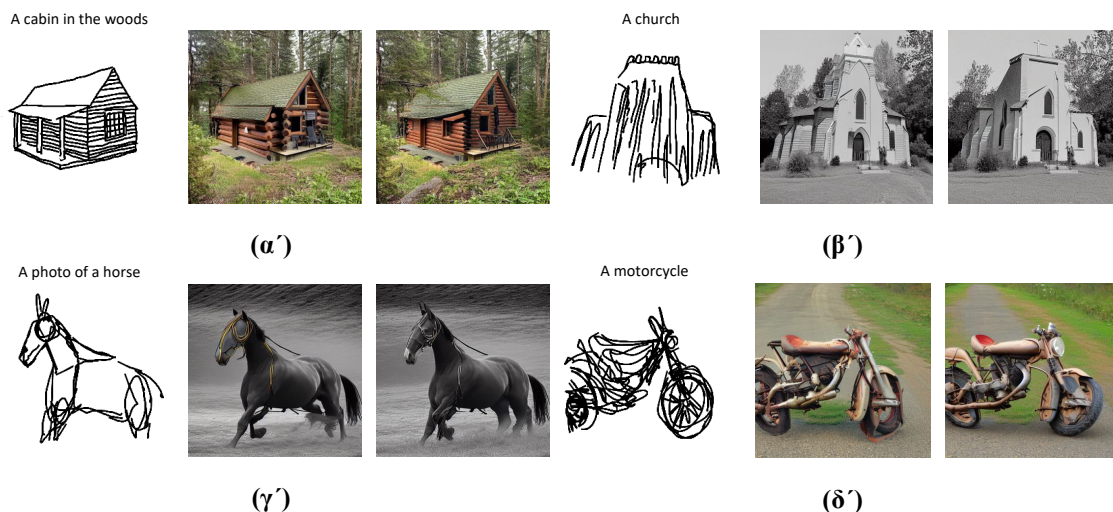
**Σχήμα 7.5:** Παραδείγματα καθοδηγούμενης από σκίτσα σύνθεσης εικόνων, με χρήση του U-Net latent edge predictor και διαφορετικές τιμές αριθμού βημάτων αποθρομβοποίησης.

η αύξηση του αριθμού των βημάτων αποθρομβοποίησης βελτιώνει εν γένει τις παραγόμενες εικόνες ως προς την πιστότητα των ακμών των αντίστοιχων σκίτσων αναφοράς, χωρίς αυτό να συνεπάγεται πάντοτε και ολική βελτίωση του ρεαλισμού και της ευκρίνειάς τους, όπως μπορεί κανείς να παρατηρήσει στην περίπτωση του ιστιοφόρου της πρώτης σειράς. Η πιο σημαντική παρατήρηση ωστόσο έγκειται στο ότι οι αλλαγές που παρουσιάζονται στις συνθετικές εικόνες όταν αυξάνεται ο αριθμός των βημάτων αποθρομβοποίησης, δεν είναι ιδιαίτερες σημαντικές, ώστε να κρίνεται απαραίτητη η αύξηση του αριθμού αυτού για την παραγωγή ικανοποιητικών αποτελεσμάτων. Το γεγονός αυτό είναι ιδιαίτερος καθοριστικό, καθώς η αύξηση του αριθμού των βημάτων της αντίστροφης διαδικασίας διάχυσης επιφέρει αυτομάτως και αντίστοιχη αύξηση του απαιτούμενου χρόνου εκτέλεσης. Αυτό υποδηλώνει, ότι η αρχιτεκτονική τύπου U-Net, όταν χρησιμοποιείται ως latent edge predictor, ενισχύει τη σθεναρότητα της διαδικασίας της σύνθεσης, μειώνοντας το πλήθος των απαιτούμενων βημάτων αποθρομβοποίησης και διατηρώντας συγχρόνως την ευκρίνεια και το ρεαλισμό στις παραγόμενες εικόνες.

## 7.2 Χρήση Δικτύου Απλοποίησης Σκίτσων

Όπως αναφέρθηκε στην ενότητα 4.4, ένα θεμελιώδες στοιχείο του προτεινόμενου πλαισίου για την αντιμετώπιση του προβλήματος της καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων είναι το δίκτυο απλοποίησης σκίτσων, το οποίο επιτρέπει στους χρήστες να απλοποιούν και να εξομαλύνουν τα σκίτσα τους. Το δίκτυο αυτό είναι εξαιρετικά χρήσιμο σε περιπτώσεις πρόχειρων και κακοσχεδιασμένων σκίτσων, όπου και χρησιμοποιείται για να μετριάσει την επίδραση των επικαλυπτόμενων και ασαφών γραμμών των σκίτσων αυτών στη διαδικασία καθοδήγησης. Προκειμένου να αξιολογηθεί η συμβολή του δικτύου στην επίδοση του προτεινόμενου πλαισίου, στο Σχ.7.6 παρουσιάζονται ορισμένα αποτελέσματα σύνθεσης, τα οποία προέκυψαν τόσο με, όσο και χωρίς τη χρήση του.

Βάσει των αποτελεσμάτων αυτών συμπεραίνεται, ότι το δίκτυο απλοποίησης σκίτσων βελτιώνει αισθητά τις παραγόμενες εικόνες ως προς τις διάφορες πτυχές της όλης διαδικασίας. Από το παράδειγμα 7.6α' παρατηρείται ότι η ομαλοποίηση που εφαρμόζεται στο σκίτσο εισόδου οδηγεί σε καλύτερη αποτύπωση της θέσης και του πλάτους των πυκνών παράλληλων γραμμών του στην παραγόμενη εικόνα. Ακολούθως, μέσω του παραδείγματος 7.6β' φαίνεται ότι η απλοποίηση του σκίτσου αναφοράς ελαττώνει την επίδραση των ενδιάμεσων αφηρημένων, πυκνών και άτακτων γραμμών, με αποτέλεσμα την καλύτερη ευθυγράμμιση της τελικής εικόνας με τα αντίστοιχα χωρικά περιγράμματα αναφοράς. Τέλος, τα παραδείγματα 7.6γ' και 7.6δ' υποδεικνύουν ότι έπειτα από την εξομάλυνση των σκίτσων εισόδου, επέρχεται μία συνολική βελτίωση στις παραγόμενες εικόνες, τόσο ως προς το ρεαλισμό όσο και ως προς την πιστότητα των χωρικών περιγραμμάτων αναφοράς.



**Σχήμα 7.6:** Παραδείγματα καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων. Για κάθε τριάδα παρουσιάζονται από αριστερά προς τα δεξιά: το σκίτσο αναφοράς και η αντίστοιχη κειμενική περιγραφή, η εικόνα που παράγεται χωρίς και με χήση του δικτύου απλοποίησης σκίτσων.

### 7.3 Γενικά συμπεράσματα

Συνοψίζοντας στο σημείο αυτό όλες τις προαναφερθείσες παρατηρήσεις και διαπιστώσεις, συμπεραίνεται ότι η χρήση του U-Net latent edge predictor βελτιώνει αισθητά τη διαδικασία της καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεση εικόνων.

Η ικανότητα της αρχιτεκτονικής να εξάγει χωρικές συσχετίσεις, οδηγεί στην παραγωγή πολύ ρεαλιστικών εικόνων, οι οποίες ακολουθούν σε πολύ ικανοποιητικό βαθμό τα αντίστοιχα σκίτσα αναφοράς. Συγχρόνως, ο αριθμός των βημάτων της αντίστροφης διαδικασίας διάχυσης που απαιτείται για την παραγωγή των αποτελεσμάτων αυτών, ανέρχεται σε μόλις 50, με άμεση συνέπεια τη σημαντική μείωση του αντίστοιχου χρόνου εκτέλεσης. Ενδεικτικά, ο χρόνος αυτός, όταν χρησιμοποιείται μία *NVIDIA T4* κάρτα γραφικών, ανέρχεται σε περίπου 50 δευτερόλεπτα ανά εικόνα. Η αρχικοποίηση του Γκαουσιανού θορύβου κατά την εκκίνηση της διαδικασίας, παρουσιάζει την εντονότερη επίδραση στα παραγόμενα αποτελέσματα. Έτσι, αρχικοποιήσεις θορύβου με ενδογενείς χωρικές διατάξεις ευθυγραμμισμένες με τα αντίστοιχα περιγράμματα των σκίτσων αναφοράς διευκολύνουν σημαντικά τη διαδικασία της σύνθεσης και οδηγούν σε εικόνες με υψηλό ρεαλισμό και πιστότητα ακμών. Αντιθέτως, αρχικοποιήσεις των οποίων οι ενδογενείς διατάξεις αποκλίνουν σημαντικά από τις αντίστοιχες γεωμετρίες των σκίτσων, έχουν ως αποτέλεσμα την παραγωγή είτε λιγότερο ρεαλιστικών, είτε αποκλινόντων ως προς τη γεωμετρία αποτελεσμάτων. Παρά ταύτα, ο U-Net latent edge predictor παρουσιάζει μία μορφή σθεναρότητας ως προς τις διαφορετικές αρχικοποιήσεις, με τις παραγόμενες εικόνες να μην είναι μεν το ίδιο ικανοποιητικές ως προς τη γενική δομική ομοιότητα με τα σκίτσα αναφοράς, συλλαμβάνοντας και αποτυπώνοντας δε το περιεχόμενο της κειμενικής περιγραφής και τα βασικά περιγράμματα των σκίτσων αυτών. Τέλος, ο βαθμός στον οποίο η καθοδήγηση λόγω σκίτσου μεταβάλει τα αποτελέσματα είναι εξέχουσας σημασίας και θα πρέπει να επιλέγεται, σε συνάρτηση πάντοτε και με το αντίστοιχο πλήθος των βημάτων αποθορυβοποίησης, ώστε να εξασφαλίζεται μία ισορροπία μεταξύ ρεαλισμού και πιστότητας ως προς τη γεωμετρία αναφοράς.

Στο επόμενο Κεφάλαιο θα γίνει μία άμεση ποιοτική και ποσοτική σύγκριση των αποτελεσμάτων της διαδικασίας της καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων, με χρήση των δύο δικτύων MLP και U-Net ως latent edge predictors, προκειμένου να εξαχθούν συμπεράσματα και να επιβεβαιωθεί η υπεροχή της προτεινόμενης αρχιτεκτονικής.

# Κεφάλαιο 8

## Συγκριτική Αξιολόγηση

---

Στα προηγούμενα κεφάλαια παρατέθηκαν παραδείγματα καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων, με χρήση των δύο αρχιτεκτονικών latent edge predictors και εξήχθησαν συμπεράσματα σχετικά με την αποτελεσματικότητα και την επίδοση καθενός εξ αυτών. Επιπλέον, μελετήθηκε ο τρόπος με τον οποίο οι επιμέρους υπερπαραμέτροι επηρεάζουν τη διαδικασία της σύνθεσης σε κάθε περίπτωση και έγινε σχολιασμός του τρόπου με τον οποίο αυτές θα πρέπει να επιλέγονται συστηματικά, ώστε να εξασφαλίζονται ικανοποιητικά αποτελέσματα. Η μελέτη αυτή πραγματοποιήθηκε μεμονωμένα για κάθε latent edge predictor, χωρίς να γίνει κάποια άμεση αντιπαράθεση των αποτελεσμάτων των δύο περιπτώσεων, κάτι το οποίο και είναι απαραίτητο, προκειμένου να διαπιστωθεί ποια από τις δύο αρχιτεκτονικές είναι πιο αποτελεσματική στην αντιμετώπιση του προβλήματος της καθοδηγούμενης από σκίτσα σύνθεσης εικόνων.

Για το λόγο αυτό, το παρόν κεφάλαιο είναι αφιερωμένο εξ ολοκλήρου στη συγκριτική αξιολόγηση της επίδοσης των αρχιτεκτονικών του per-pixel MLP και του U-Net latent edge predictor και αποσκοπεί στην εξαγωγή συμπερασμάτων σχετικά την υπεροχή κάποιου εξ αυτών. Στις επόμενες ενότητες θα παρουσιαστούν αντιπαραθετικά, αρχικά ποιοτικά και έπειτα ποσοτικά αποτελέσματα, ενώ στο τέλος θα γίνει ένας γενικός σχολιασμός και θα υπογραμμιστούν ορισμένα βασικά συμπεράσματα και παρατηρήσεις. Στο σημείο αυτό, σημειώνεται για ακόμη μία φορά, ότι η αρχιτεκτονική του per-pixel MLP είναι αυτή που χρησιμοποιείται ως latent edge predictor στην πρωτότυπη υλοποίηση των *Voynov et al.* [4], ενώ η αρχιτεκτονική τύπου U-Net είναι η αρχιτεκτονική που προτείνεται στην παρούσα εργασία.

### 8.1 Ποιοτική Αξιολόγηση

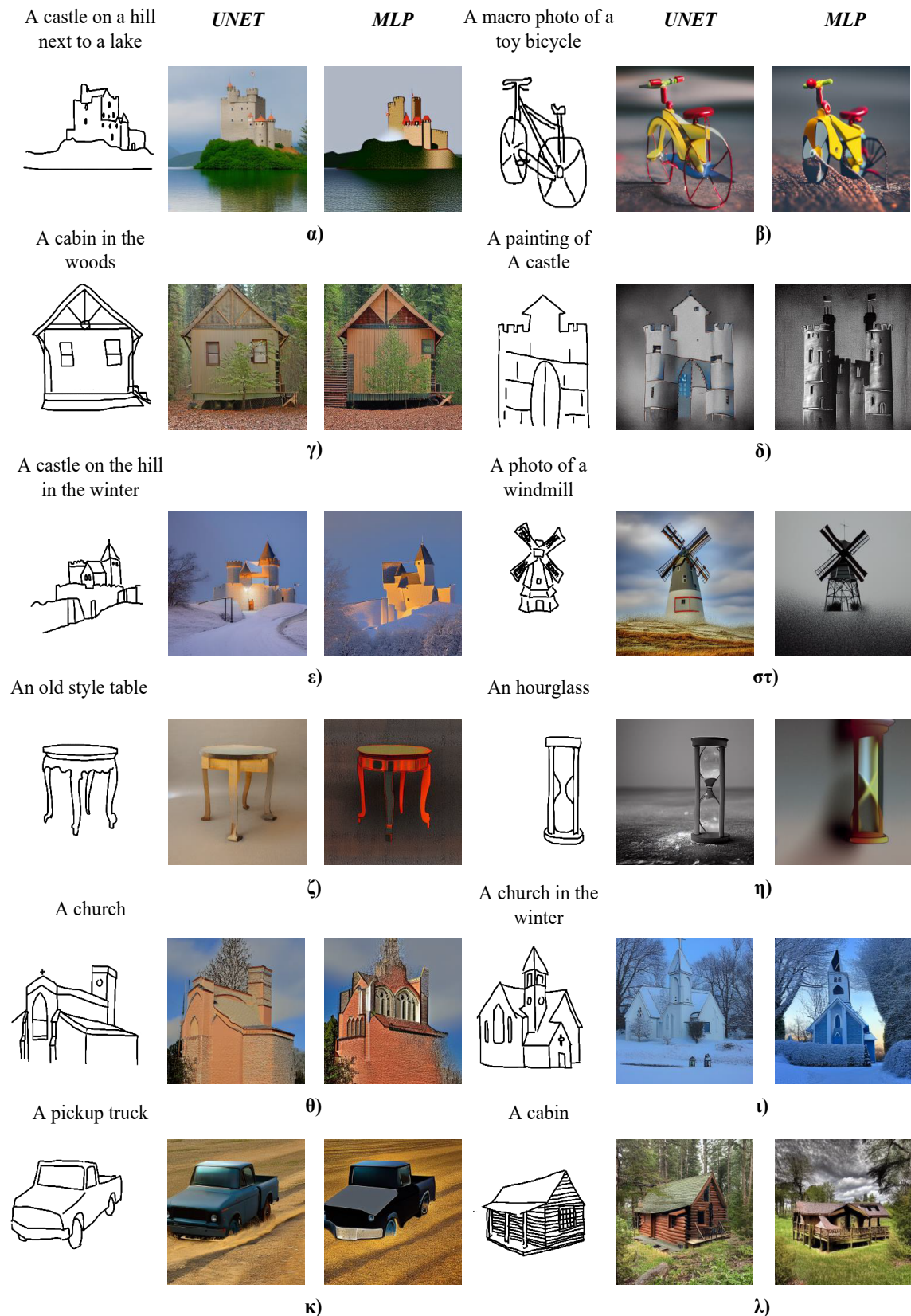
Η πρώτη ενότητα του παρόντος κεφαλαίου εστιάζεται στην ποιοτική σύγκριση και αξιολόγηση των αποτελεσμάτων των δύο προαναφερθέντων αρχιτεκτονικών, ήτοι στην οπτική αξιολόγηση συνθετικών εικόνων που προκύπτουν με χρήση των αρχιτεκτονικών αυτών ως latent edge predictors. Στα Σχ.8.1 και 8.2 παρατίθενται παραδείγματα σύνθεσης με χρήση των δύο δικτύων. Όσον αφορά στις τιμές των παραμέτρων που χρησιμοποιούνται στη διαδικασία της σύνθεσης, επιλέγονται αυτές που παρουσιάζονται στον Πίνακα 7.2. Στο πλαίσιο αυτό, ειδική αναφορά αξίζει να γίνει στον αριθμό των βημάτων αποθορυβο-



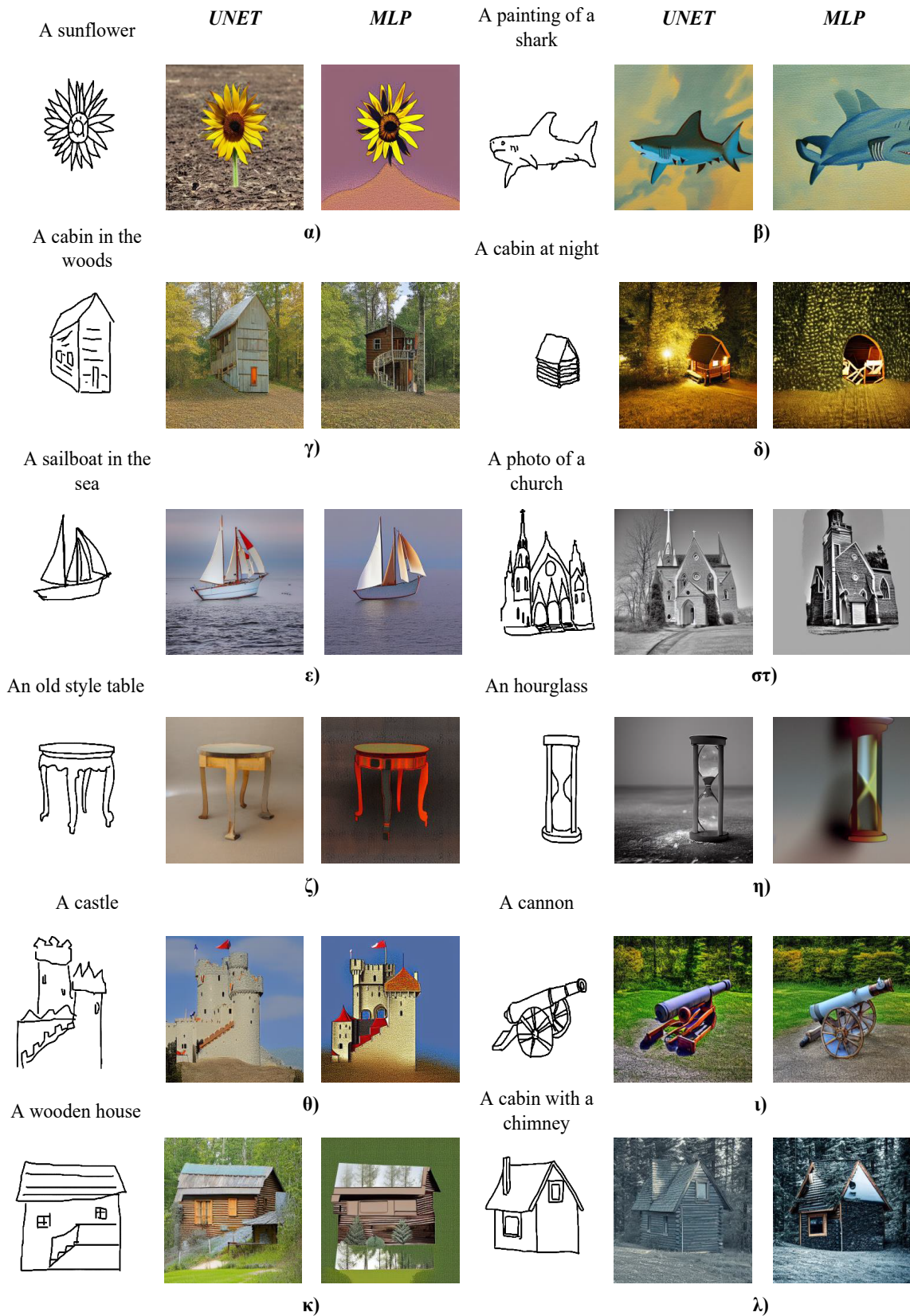
ποίησης. Σκοπός κάθε υλοποίησης είναι η μείωση του χρονικού και υπολογιστικού κόστους, χωρίς ωστόσο η μείωση αυτή να συντελείται σε βάρος της ποιότητας των αποτελεσμάτων. Στο Κεφάλαιο 7 διαπιστώθηκε ότι ο αριθμός των 50 βημάτων αποθορυβοποίησης είναι αρκετός, στην περίπτωση του U-Net latent edge predictor, ώστε να οδηγήσει στην παραγωγή ρεαλιστικών και πιστών ως προς τη γεωμετρία των σκίτσων εικόνων. Είναι επομένως θεμιτό, η σύγκριση των δύο αρχιτεκτονικών να γίνει στη βάση του ελάχιστου αριθμού βημάτων, ο οποίος εγγυάται ικανοποιητικά αποτελέσματα και κατά συνέπεια στη βάση του ελάχιστου χρονικού κόστους. Ως εκ τούτου, ο αριθμός των βημάτων αποθορυβοποίησης που χρησιμοποιείται σε όλα τα παρακάτω παραδείγματα, ισούται με 50. Τέλος, σημειώνεται, ότι προκειμένου η σύγκριση των δύο αρχιτεκτονικών να είναι όσο το δυνατόν πιο αντικειμενική και αμερόληπτη, σε κάθε τριάδα παραδειγμάτων, χρησιμοποιείται η ίδια αρχικοποίηση θορύβου, η οποία όπως διαπιστώθηκε και σε προηγούμενες ενότητες, επηρεάζει σε πολύ μεγάλο βαθμό τα τελικά αποτελέσματα.

Βάσει τώρα των παραδειγμάτων των Σχ.8.1 και 8.2 διαπιστώνονται τα εξής:

- Η χρήση του U-Net ως latent edge predictor παρουσιάζει μία συνολική υπεροχή έναντι του per-pixel MLP, όσον αφορά στο ρεαλισμό και την ευκρίνεια των παραγόμενων εικόνων. Πιο συγκεκριμένα, η χρήση του U-Net οδηγεί στην παραγωγή εικόνων, των οποίων τα εικονιζόμενα αντικείμενα ενδιαφέροντος χαρακτηρίζονται από υψηλό βαθμό ρεαλισμού και φυσικότητας, σε αντίθεση με τις αντίστοιχες εικόνες του per-pixel MLP, οι οποίες υστερούν στη γενική περίπτωση, τόσο σε ρεαλισμό όσο και σε ευκρίνεια. Στο πλαίσιο αυτό, χαρακτηριστικά είναι τα παραδείγματα α), γ), ε), ζ), η) του Σχ.8.1 και τα α), ε), ζ) και στ) του Σχ.8.2. Στα παραδείγματα αυτά, οι εικόνες που προκύπτουν με χρήση του per-pixel MLP, είτε αποκλίνουν σημαντικά από την πτυχή του ρεαλισμού (anime-like μορφή), είτε διατηρούν το ρεαλισμό τους, χωρίς όμως να παρουσιάζουν την επιθυμητή ευκρίνεια. Αντιθέτως, οι ελλείψεις αυτές δεν παρατηρούνται στις αντίστοιχες εικόνες του U-Net, οι οποίες πέραν του στοιχείου του ρεαλισμού, επιδεικνύουν και αισθητά καλύτερη ευκρίνεια.
- Η χρήση του U-Net παρουσιάζει αντίστοιχη υπεροχή και ως προς την αποτύπωση των περιγραμμάτων των σκίτσων αναφοράς. Χαρακτηριστικά είναι τα παραδείγματα β), δ), θ) και λ) του Σχ.8.1, καθώς και τα παραδείγματα β), γ), δ), ι) και λ) του Σχ.8.2. Στα παραδείγματα αυτά, οι εικόνες που παράγονται με χρήση του per-pixel MLP latent edge predictor, αδυνατούν να αποτυπώσουν τη γεωμετρική διάταξη των σκίτσων αναφοράς, αποκλίνοντας είτε σε μικρό βαθμό (π.χ. β), δ) Σχ.8.1), είτε σε πολύ σημαντικό βαθμό από αυτή (π.χ. θ), ι) λ) Σχ.8.1 και παραδείγματα β), γ), δ) και ι) Σχ.8.2).
- Σε περιπτώσεις όπου τα περιγράμματα των σκίτσων είναι ιδιαίτερος περίπλοκα, ο U-Net latent edge predictor τα αποτυπώνει με πιο ελεύθερο τρόπο, προκειμένου να διατηρηθεί ο συνολικός ρεαλισμός των παραγόμενων εικόνων (π.χ. θ), ι) Σχ.8.1 και η), θ) κ) Σχ.8.2), κάτι το οποίο δε συμβαίνει και στην περίπτωση του per-pixel MLP, όπου οι παραγόμενες εικόνες, είτε ακολουθούν πιστά τα περιγράμματα των σκίτσων, χάνοντας σε πολύ μεγάλο βαθμό το ρεαλισμό τους (π.χ. ε), κ) Σχ.8.1 και α) η) Σχ.8.2), είτε αδυνατούν πλήρως να αποτυπώσουν την πολύπλοκη αυτή γεωμετρία (π.χ. δ), κ) Σχ.8.2).



**Σχήμα 8.1:** Παραδείγματα καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων με χρήση αμφοτέρων των U-Net και per-pixel MLP latent edge predictors. Σε κάθε τριάδα παραδειγμάτων παρατίθενται από αριστερά προς τα δεξιά: το σκίτσο αναφοράς και η κειμενική περιγραφή, η παραγόμενη εικόνα με χρήση του U-Net και η παραγόμενη εικόνα με χρήση του per-pixel MLP latent edge predictor, για  $T = 50$  βήματα αποθρομβοποίησης.



**Σχήμα 8.2:** Παραδείγματα καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων με χρήση αμφοτέρων των U-Net και per-pixel MLP latent edge predictors. (συνέχεια)

Συνολικά, η χρήση του U-Net latent edge predictor βελτιώνει αισθητά τα αποτελέσματα της διαδικασίας της καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων, κατορθώνοντας σε 50 μόλις βήματα αποθρομβοποίησης (υποπενταπλάσιος αριθμός από αυτόν της περίπτωσης του per-pixel MLP) να παράξει εικόνες, οι οποίες χαρακτηρίζονται από υψηλό επίπεδο ρεαλισμού, ευκρίνειας και πιστότητας ως προς τις ακμές των σκίτσων αναφοράς. Συγχρόνως, ιδιαίτερος ικανοποιητική είναι στην περίπτωση αυτή και η συμμόρφωση των παραγόμενων εικόνων με τις αντίστοιχες κειμενικές περιγραφές. Αντιθέτως, η χρήση του per-pixel MLP latent edge predictor δεν είναι εξίσου αποδοτική στην περίπτωση των 50 βημάτων αποθρομβοποίησης, με τις παραγόμενες εικόνες να υστερούν, είτε ως προς το ρεαλισμό και την ευκρίνεια, είτε ως προς την απόδοση της γεωμετρικής διάταξης των σκίτσων αναφοράς, είτε και ως προς τις δύο αυτές συνιστώσες. Προφανώς, αύξηση του αριθμού των βημάτων αποθρομβοποίησης, θα έχει ως αποτέλεσμα και αντίστοιχη βελτίωση της επίδοσης του per-pixel MLP. Η ενέργεια αυτή ωστόσο είναι μη θεμιτή και ασύμφορη, καθώς συνεπάγεται σημαντική αύξηση του χρονικού κόστους. Τέλος, ιδιαίτερος σημαντική είναι και η συνεισφορά του U-Net latent edge predictor στην ενίσχυση της σθεναρότητας της διαδικασίας της σύνθεσης, υπό την έννοια ότι ακόμη και στην περίπτωση όπου τα παραγόμενα αποτελέσματα αποκλίνουν σε κάποιο βαθμό από τις αντίστοιχες περίπλοκες γεωμετρικές αναφορές, εξισορροπούν την απώλεια αυτή, μέσω της διατήρησης της φυσικότητας και του ρεαλισμού τους.

## 8.2 Ποσοτική Αξιολόγηση

Έως το σημείο αυτό, η αξιολόγηση των αποτελεσμάτων των δύο αρχιτεκτονικών των latent edge predictors ήταν, ως επί το πλείστον, ποιοτική. Στην παρούσα ενότητα θα γίνει μία προσπάθεια ποσοτικοποίησης της επίδοσης των δύο αυτών αρχιτεκτονικών, ούτως ώστε να επιβεβαιωθεί η υπεροχή του προτεινόμενου U-Net latent edge predictor έναντι του αντίστοιχου per-pixel MLP.

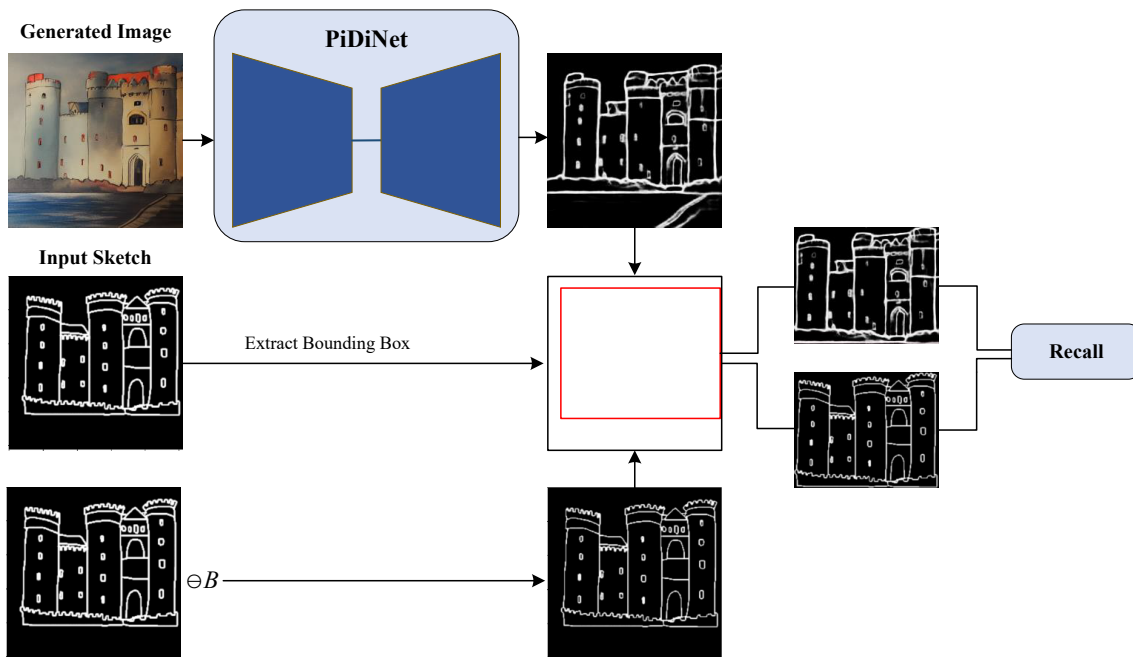
Μια σημαντικότερη πρόκληση του προβλήματος της καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων είναι η απουσία μετρικών, ικανών να συνυπολογίζουν με αποτελεσματικό τρόπο, τις πτυχές της πιστότητας ως προς την απόδοση των ακμών, του ρεαλισμού και της συνολικής ποιότητας των παραγόμενων εικόνων. Προκειμένου να αντιμετωπιστεί η δυσκολία αυτή, στα πλαίσια της παρούσας εργασίας πραγματοποιούνται δύο *User Studies* και βάσει της διαμορφούμενης κοινής γνώμης, γίνεται ποσοτικοποίηση της επίδοσης των δύο επιμέρους latent edge predictors. Επιπλέον και προκειμένου να αποκτηθεί μια στοιχειώδης και πρωτόλεια ποσοτική εικόνα της ικανότητας των δύο αρχιτεκτονικών για απόδοση των ακμών των σκίτσων αναφοράς, γίνεται χρήση της μετρική *recall*.

### 8.2.1 Recall

Στα πλαίσια της παρούσας ποσοτικής αξιολόγησης, η μετρική *recall* χρησιμοποιείται για τη σύγκριση των σκίτσων αναφοράς με τους χάρτες ακμών, οι οποίοι εξάγονται από τις αντίστοιχες παραγόμενες εικόνες. Για την εξαγωγή των χαρτών αυτών αξιοποιείται και πάλι το δίκτυο PiDiNet, καθώς έχει τη δυνατότητα να αγνοεί ορισμένες παρασκηνιακές

και ιδιαίτερες περίπλοκες ακμές, οι οποίες και δεν παρουσιάζουν ενδιαφέρον (π.χ. σύνθετες ακμές που εντοπίζονται στο φόντο). Δεδομένου ότι σκοπός είναι η σύγκριση να περιοριστεί στις περιοχές, οι οποίες περιέχουν τα αντικείμενα ενδιαφέροντος και λαμβάνοντας υπόψιν ότι οι συνθετικές εικόνες παρουσιάζουν πολύ υψηλό επίπεδο πολυπλοκότητας όσον αφορά στα εικονιζόμενα τοπία, η χρήση ανιχνευτών ακμών όπως ο Canny Edge Detector ή το φίλτρο Sobel, οδηγεί σε χάρτες ακμών με αναλογικά ιδιαίτερος υψηλή πολυπλοκότητα. Ο πλεονασμός στις ανιχνευθείσες ακμές που επιφέρει η χρήση τέτοιου είδους ανιχνευτών, οδηγεί σε πολύ χαμηλές τιμές της μετρικής του recall, λόγω του μεγάλου αριθμού των ψευδώς θετικών (false positive) pixels.

Πριν γίνει η σύγκριση των σκίτσων αναφοράς με τους αντίστοιχους εξαγόμενους χάρτες ακμών, κρίνεται σκόπιμο να υλοποιηθεί ένα επιπρόσθετο βήμα προεπεξεργασίας. Πιο συγκεκριμένα, λόγω του ελεύθερου και ατημέλητου χαρακτήρα τους, οι γραμμές που απαρτίζουν τα σκίτσα εισόδου παρουσιάζουν μεταβαλλόμενο πάχος, το οποίο ενδέχεται να μην είναι κατάλληλο για άμεση σύγκριση με τις αντίστοιχες εξαγόμενες ακμές. Προκειμένου να αμβλυνθεί το πρόβλημα αυτό, εφαρμόζεται στα σκίτσα εισόδου ο μορφολογικός τελεστής του *erosion*, χρησιμοποιώντας ως δομικό στοιχείο  $B$  ένα all-ones  $3 \times 3$  μητρώο. Τα eroded πλέον σκίτσα, αξιολογούνται εν συνεχεία ως αναφορά για τον υπολογισμό της μετρικής του recall. Τέλος, βάσει των περιγραμμάτων των εικονιζόμενων αντικειμένων, εξάγονται κατάλληλα *bounding boxes*, τα οποία χρησιμοποιούνται για την περικοπή τόσο των σκίτσων αναφοράς, όσο και των χαρτών ακμών των παραγόμενων εικόνων. Η περικοπή αυτή περιορίζει σημαντικά τη σύγκριση στις περιοχές ενδιαφέροντος. Η όλη διαδικασία για τον υπολογισμό της μετρικής του recall αποδίδεται σχηματικά στο Σχ.8.3.



**Σχήμα 8.3:** Διαδικασία υπολογισμού μετρικής Recall μεταξύ σκίτσου αναφοράς και εξαγόμενου χάρτη ακμών της παραγόμενης εικόνας.

Για τον υπολογισμό της μετρικής του recall σύμφωνα με την ανωτέρω διαδικασία, παράχθηκαν συνολικά 1400 εικόνες. Οι αριθμητικές τιμές που προέκυψαν για την περίπτωση του MLP και του προτεινόμενου U-Net latent edge predictor παρουσιάζονται συγκεντρωτικά στον Πίνακα 8.1.

### 8.2.2 User Studies

Όπως αναφέρθηκε και ανωτέρω, βασικός πυλώνας της ποσοτικής αξιολόγησης της επίδοσης των δύο αρχιτεκτονικών καθοδήγησης, είναι η γνώμη των χρηστών, η οποία και διαμορφώνεται μέσα από κατάλληλα user studies. Στα πλαίσια της παρούσας εργασίας, διεξήχθησαν δύο ολοκληρωμένα και πλήρη user studies, υπό τη μορφή ερωτηματολογίων. Τα ερωτηματολόγια αυτά συμπληρώθηκαν και υποβλήθηκαν ανώνυμα, με τους συμμετέχοντες να μη γνωρίζουν την εκάστοτε υποκείμενη αρχιτεκτονική που χρησιμοποιήθηκε για την παραγωγή των αντίστοιχων εικόνων, έτσι ώστε να εξασφαλιστεί η αντικειμενικότητα και η αμεροληψία της όλης διαδικασίας.

Στην πρώτη εκ των δύο μελετών, η οποία από τούδε και στο εξής θα καλείται *User Preference Evaluation Study*, συμμετείχαν 37 άτομα, στα οποία παρουσιάστηκαν σκίτσα συνοδευόμενα από κειμενικές περιγραφές, μαζί με τις αντίστοιχες εικόνες που παρήχθησαν με χρήση του MLP και U-Net latent edge predictor. Για κάθε σκίτσο και κειμενική περιγραφή εισόδου, χρησιμοποιήθηκε η ίδια αρχικοποίηση θορύβου, έτσι ώστε να εξασφαλιστεί η αμεροληψία ως προς τα αποτελέσματα των δύο αρχιτεκτονικών. Η επιλογή των παραδειγμάτων πραγματοποιήθηκε με τυχαίο τρόπο, με τους συμμετέχοντες να μην διαθέτουν γνώση σχετικά με το ποια αρχιτεκτονική παρήγαγε κάθε εικόνα. Δοθέντων των παραδειγμάτων αυτών, οι χρήστες κλήθηκαν να αξιολογήσουν κάθε παράδειγμα σύνθεσης, με τρόπο ώστε να υποδείξουν ποια από τις παραγόμενες εικόνες (α) είναι πιο ρεαλιστική, (β) είναι πιο αποδοτική ως προς την πιστότητα των ακμών και (γ) παρουσιάζει μεγαλύτερη γενική δομική ομοιότητα συγκριτικά με το σκίτσο αναφοράς. Συνολικά, συλλέχθησαν 1110 συγκριτικές αξιολογήσεις, οι οποίες και επιβεβαίωσαν την υπεροχή του προτεινόμενου U-Net latent edge predictor, έναντι του αντίστοιχου MLP, όπως φαίνεται από τα αποτελέσματα του Πίνακα 8.1. Πιο συγκεκριμένα, οι εικόνες που προέκυψαν από την προτεινόμενη αρχιτεκτονική, κρίθηκαν πιο ρεαλιστικές από το 60.9% της κοινής γνώμης και πιο πιστές ως προς τα περιγράμματα των σκίτσων αναφοράς, από το 70.5% της κοινής γνώμης. Τέλος, όσον αφορά στη γενική δομική συνοχή, το 70.7% των χρηστών έκρινε ότι η προτεινόμενη μέθοδος υπερτερεί της αντίστοιχη μεθόδου η οποία βασίζεται στη χρήση του MLP latent edge predictor. Συνολικά, τα αποτελέσματα της πρώτης αυτής μελέτης επιβεβαιώνουν και ενισχύουν τα ήδη εξαγόμενα συμπεράσματα σχετικά με την υπεροχή της προτεινόμενης U-Net αρχιτεκτονικής, έναντι της αντίστοιχης MLP αρχιτεκτονικής.

Στη δεύτερη μελέτη, η οποία αναφέρεται ως *User Rating Evaluation Study*, παρουσιάστηκε σε μία ομάδα 31 συμμετεχόντων ένα σύνολο σκίτσων μαζί με τις αντίστοιχες κειμενικές περιγραφές. Κάθε σκίτσο και κειμενική περιγραφή συνοδευόταν από την αντίστοιχη παραγόμενη εικόνα, η οποία παράχθηκε είτε με χρήση του per-pixel MLP, είτε του προτεινό-

**Πίνακας 8.1:** Αποτελέσματα ποσοτικής αξιολόγησης.

| Latent Edge Predictor | Recall (↑)   | Human Preference Evaluation Study (↑) |                 |                         | Human Rating Evaluation Study (↑) |                                   |   |
|-----------------------|--------------|---------------------------------------|-----------------|-------------------------|-----------------------------------|-----------------------------------|---|
|                       |              | Ρεαλισμός                             | Πιστότητα ακμών | Γενική δομική ομοιότητα | Ρεαλισμός ( $\pm\sigma$ )         | Πιστότητα ακμών ( $\pm\sigma$ )   | Γενική δομική ομοιότητα ( $\pm\sigma$ ) |
| MLP <sub>LEP</sub>    | 0.595        | 39.1%                                 | 29.5%           | 29.3%                   | 3.29 $\pm$ 1.04                   | 2.89 $\pm$ 1.02                   | 3.20 $\pm$ 1                            |
| U-Sketch              | <b>0.645</b> | <b>60.9%</b>                          | <b>70.5%</b>    | <b>70.7%</b>            | <b>3.97 <math>\pm</math> 1</b>    | <b>3.86 <math>\pm</math> 0.95</b> | <b>4.03 <math>\pm</math> 0.93</b>       |

μενου U-Net latent edge predictor, γεγονός το οποίο αγνοούσαν πλήρως οι συμμετέχοντες, ώστε να εξασφαλίζεται το αδιάβλητο της όλης διαδικασίας. Βάσει των παρατιθέμενων αυτών παραδειγμάτων, οι χρήστες κλήθηκαν να αξιολογήσουν τις παραγόμενες εικόνες σε κλίμακα από το 1 έως το 5 (κακή  $\rightarrow$  άριστη) ως προς (α) το ρεαλισμό, (β) την πιστότητα των ακμών και (γ) τη γενική δομική συνοχή τους. Συνολικά συγκεντρώθηκαν 930 αξιολογήσεις. Από τις συνολικές παρατιθέμενες εικόνες, οι μισές παρήχθησαν με χρήση του MLP και οι υπόλοιπες μισές με χρήση του U-Net latent edge predictor. Όπως και προηγουμένως, τα αποτελέσματα της μελέτης αυτής παρουσιάζονται συγκεντρωτικά στον Πίνακα 8.1. Κάθε καταχώρηση του πίνακα αντιστοιχεί στη μέση βαθμολογία κοινής γνώμης, η οποία υπολογίζεται βάσει της μέσης τιμής ( $\mu$ ) και της τυπικής απόκλισης ( $\sigma$ ) των απαντήσεων των χρηστών. Από τα αποτελέσματα αυτά διαπιστώνεται και πάλι, ότι η προτεινόμενη αρχιτεκτονική υπερτερεί έναντι του αντίστοιχου MLP latent edge predictor σε όλες τις επιμέρους πτυχές που εξετάζονται. Πιο αναλυτικά, η χρήση του U-Net latent edge predictor επιφέρει βελτίωση της τάξεως του 20.4% ως προς το ρεαλισμό, 33.8% ως προς την πιστότητα των ακμών και 26.1% ως προς τη γενική δομική συνοχή των παραγόμενων εικόνων σε σχέση με τα αντίστοιχα σκίτσα αναφοράς. Τέλος, ιδιαίτερος σημαντική είναι η ελάττωση της τυπικής απόκλισης που παρατηρείται στην περίπτωση των μετρικών της προτεινόμενης αρχιτεκτονικής, η οποία υποδηλώνει αύξηση στη συναίνεση και τη συμφωνία μεταξύ των συμμετεχόντων, όσον αφορά στην ποιότητα των αποτελεσμάτων.

Συνοψίζοντας, τα ποσοτικά αποτελέσματα του Πίνακα 8.1 συνηγορούν υπέρ της προτεινόμενης αρχιτεκτονικής, ενισχύοντας κατά αυτό τον τρόπο τα συμπεράσματα της ενότητας 8.1. Η άμεση ποσοτικοποίηση της επίδοσης ως προς την πιστότητα των ακμών (μετρική recall), σε συνδυασμό με τις γνώμες των χρηστών, φανερώνουν μία σημαντική βελτίωση στην ποιότητα των παραγόμενων εικόνων και επιβεβαιώνουν την ικανότητα του προτεινόμενου U-Net latent edge predictor στην αντιμετώπιση του προβλήματος της καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων.

## Κεφάλαιο 9

### Συμπεράσματα και Μελλοντική Έρευνα

---

Στα πλαίσια της παρούσας εργασίας μελετάται το πρόβλημα της καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων, με χρήση μοντέλων διάχυσης. Στον πυρήνα του προτεινόμενου πλαισίου καθοδήγησης βρίσκεται το προεκπαιδευμένο μοντέλο Stable Diffusion, το οποίο και αξιοποιείται για την παραγωγή ρεαλιστικών και υψηλής ευκρίνειας εικόνων βάσει αντίστοιχων κειμενικών περιγραφών. Για την καθοδήγηση των χωρικών περιγραμμάτων των παραγόμενων εικόνων υλοποιείται ένας latent edge predictor, ο οποίος, βάσει των ενεργοποιήσεων των ενδιάμεσων στρωμάτων του δικτύου αποθρομβοποίησης του Stable Diffusion, παρέχει εκτίμηση για το χωρικό χάρτη της συντιθέμενης εικόνας σε κάθε βήμα της αντίστροφης διαδικασίας διάχυσης. Ο χάρτης αυτός αξιοποιείται εν συνεχεία για σύγκριση με το σκίτσο αναφοράς και καθοδήγηση των χωρικών περιγραμμάτων της εικόνας. Όσον αφορά στην αρχιτεκτονική του δικτύου του latent edge predictor, εξετάζονται αντιπαραθετικά η αρχιτεκτονική ενός per-pixel MLP, η οποία και υλοποιήθηκε από τους *Voynov et al.* [4] και η προτεινόμενη U-Net αρχιτεκτονική, η οποία αποσκοπεί στην εξαγωγή και ενθυλάκωση χωρικών συσχετίσεων των pixels των εισόδων. Τέλος, ο προτεινόμενο πλαίσιο επεκτείνεται με την προσθήκη ενός δικτύου απλοποίησης σκίτσων, το οποίο εξομαλύνει και απλοποιεί τα σκίτσα εισόδου και η χρήση του οποίου επαφίεται στην επιλογή του χρήστη.

Τα ποιοτικά αποτελέσματα, σε συνδυασμό με τη γνώμη των χρηστών, αποδεικνύουν την υπεροχή του προτεινόμενου U-Net latent edge predictor, ως προς τη σύνθεση ρεαλιστικών εικόνων, ικανών να αποδίδουν τα χωρικά περιγράμματα των σκίτσων αναφοράς και να συμμορφώνονται με τις αντίστοιχες κειμενικές περιγραφές. Ιδιαίτερως σημαντική κρίνεται η συνεισφορά της προτεινόμενης αρχιτεκτονικής στη μείωση του πλήθους των βημάτων, που απαιτούνται κατά τη διαδικασία αποθρομβοποίησης για την παραγωγή ικανοποιητικών αποτελεσμάτων. Στο πλαίσιο αυτό, το απαιτούμενο πλήθος των βημάτων μειώνεται από 250, της περίπτωσης του MLP latent edge predictor, σε μόλις 50 βήματα, με χρήση της προτεινόμενης U-Net αρχιτεκτονικής. Η μείωση αυτή οδηγεί σε αναλογική ελάττωση του αντίστοιχου συνολικού χρόνου εκτέλεσης κατά ~ 80%, με αποτέλεσμα η προτεινόμενη μέθοδος να καθίσταται ιδιαίτερως αποδοτική και αποτελεσματική για την αντιμετώπιση του προβλήματος της καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων.

Όσον αφορά τώρα στη μελλοντική έρευνα, επιτακτική κρίνεται αρχικά η ανάγκη για ανάπτυξη ποσοτικών μετρικών, ικανών να αποτυπώνουν και να συνδυάζουν τις πτυχές του



ρεαλισμού, της πιστότητας των χωρικών περιγραμμάτων των σκίτσων αναφοράς και της συμμόρφωσης με το νοηματικό περιεχόμενο των κειμενικών περιγραφών. Ακολούθως και έχοντας διαπιστώσει την καθοριστική επίδραση της αρχικοποίησης του θορύβου στην όλη διαδικασία, ιδιαίτερο πεδίο έρευνας μπορεί να αποτελέσει η ανάπτυξη μηχανισμών καθοδήγησης της αρχικοποίησης του θορύβου, έτσι ώστε η ενδογενής χωρική του διάταξη να ευθυγραμμίζεται κατά το δυνατόν με τα αντίστοιχα σκίτσα αναφοράς. Τέλος, μία πιθανή κατεύθυνση της ερευνητικής δραστηριότητας είναι η μελέτη και ο σχεδιασμός νέων αρχιτεκτονικών latent edge predictors, οι οποίες θα προσεγγίζουν με καλύτερο τρόπο τους χάρτες ακμών των παραγόμενων εικόνων.

# Μέρος **IV**

## Παραρτήματα

---

# Παράρτημα A

## Επιπλέον Παραδείγματα Σύνθεσης



**Σχήμα A.1:** Παραδείγματα καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων με χρήση αμφοτέρων των U-Net και per-pixel MLP latent edge predictors ( $T = 50$ ). Σε κάθε τριάδα παρουσιάζονται από αριστερά προς τα δεξιά: σκίτσο αναφοράς, παραγόμενη εικόνα με χρήση του U-Net και παραγόμενη εικόνα με χρήση του MLP latent edge predictor.



**Σχήμα Α.2:** Παραδείγματα καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων με χρήση αμφοτέρων των U-Net και per-pixel MLP latent edge predictors, ( $T = 50$ ). (συνέχεια)



**Σχήμα Α.3:** Παραδείγματα καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων με χρήση αμφοτέρων των U-Net και per-pixel MLP latent edge predictors, ( $T = 50$ ). (συνέχεια)



**Σχήμα Α.4:** Παραδείγματα καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων με χρήση αμφοτέρων των U-Net και per-pixel MLP latent edge predictors, ( $T = 50$ ). (συνέχεια)



**Σχήμα Α.5:** Παραδείγματα καθοδηγούμενης από σκίτσα και κειμενικές περιγραφές σύνθεσης εικόνων με χρήση αμφοτέρων των U-Net και per-pixel MLP latent edge predictors, ( $T = 50$ ). (συνέχεια)

## Βιβλιογραφία

---

- [1] Wengling Chen and James Hays. “SketchyGAN: Towards Diverse and Realistic Sketch to Image Synthesis”. In: *CoRR* abs/1801.02753 (2018). arXiv: 1801.02753. URL: <http://arxiv.org/abs/1801.02753>.
- [2] Xun Huang et al. “Multimodal Unsupervised Image-to-Image Translation”. In: *CoRR* abs/1804.04732 (2018). arXiv: 1804.04732. URL: <http://arxiv.org/abs/1804.04732>.
- [3] Sheng-Yu Wang, David Bau, and Jun-Yan Zhu. “Sketch Your Own GAN”. In: *CoRR* abs/2108.02774 (2021). arXiv: 2108.02774. URL: <https://arxiv.org/abs/2108.02774>.
- [4] Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. *Sketch-Guided Text-to-Image Diffusion Models*. 2022. arXiv: 2211.13752 [cs.CV].
- [5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. *Denosing Diffusion Probabilistic Models*. 2020. arXiv: 2006.11239 [cs.LG].
- [6] Yang Song and Stefano Ermon. *Generative Modeling by Estimating Gradients of the Data Distribution*. 2020. arXiv: 1907.05600 [cs.LG].
- [7] Yang Song and Stefano Ermon. *Improved Techniques for Training Score-Based Generative Models*. 2020. arXiv: 2006.09011 [cs.LG].
- [8] Yang Song et al. *Score-Based Generative Modeling through Stochastic Differential Equations*. 2021. arXiv: 2011.13456 [cs.LG].
- [9] Aapo Hyvärinen. “Estimation of Non-Normalized Statistical Models by Score Matching”. In: *J. Mach. Learn. Res.* 6 (2005), pp. 695–709.
- [10] Yang Song and Diederik P. Kingma. *How to Train Your Energy-Based Models*. 2021. arXiv: 2101.03288 [cs.LG].
- [11] Yang Song et al. “Sliced Score Matching: A Scalable Approach to Density and Score Estimation”. In: *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*. Ed. by Ryan P. Adams and Vibhav Gogate. Vol. 115. Proceedings of Machine Learning Research. PMLR, 2020, pp. 574–584. URL: <https://proceedings.mlr.press/v115/song20a.html>.
- [12] Pascal Vincent. “A Connection Between Score Matching and Denosing Autoencoders”. In: *Neural Computation* 23.7 (2011), pp. 1661–1674. DOI: 10.1162/NECO\_a\_00142.
- [13] Max Welling and Yee Whye Teh. “Bayesian Learning via Stochastic Gradient Langevin Dynamics”. In: *ICML’11*. Bellevue, Washington, USA: Omnipress, 2011, pp. 681–688. ISBN: 9781450306195.



- [14] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi. “Optimization by Simulated Annealing”. In: *Readings in Computer Vision*. Ed. by Martin A. Fischler and Oscar Firschein. San Francisco (CA): Morgan Kaufmann, 1987, pp. 606–615. ISBN: 978-0-08-051581-6. DOI: <https://doi.org/10.1016/B978-0-08-051581-6.50059-3>. URL: <https://www.sciencedirect.com/science/article/pii/B9780080515816500593>.
- [15] Radford M. Neal. *Annealed Importance Sampling*. 1998. arXiv: [physics/9803008](https://arxiv.org/abs/physics/9803008) [[physics.comp-ph](https://arxiv.org/abs/physics/9803008)].
- [16] Brian D.O. Anderson. “Reverse-time diffusion equation models”. In: *Stochastic Processes and their Applications* 12.3 (1982), pp. 313–326. ISSN: 0304-4149. DOI: [https://doi.org/10.1016/0304-4149\(82\)90051-5](https://doi.org/10.1016/0304-4149(82)90051-5). URL: <https://www.sciencedirect.com/science/article/pii/0304414982900515>.
- [17] Jiaming Song, Chenlin Meng, and Stefano Ermon. *Denoising Diffusion Implicit Models*. 2022. arXiv: [2010.02502](https://arxiv.org/abs/2010.02502) [[cs.LG](https://arxiv.org/abs/2010.02502)].
- [18] Shakir Mohamed and Balaji Lakshminarayanan. *Learning in Implicit Generative Models*. 2017. arXiv: [1610.03483](https://arxiv.org/abs/1610.03483) [[stat.ML](https://arxiv.org/abs/1610.03483)].
- [19] Andrew Brock, Jeff Donahue, and Karen Simonyan. *Large Scale GAN Training for High Fidelity Natural Image Synthesis*. 2019. arXiv: [1809.11096](https://arxiv.org/abs/1809.11096) [[cs.LG](https://arxiv.org/abs/1809.11096)].
- [20] Ian J. Goodfellow et al. *Generative Adversarial Networks*. 2014. arXiv: [1406.2661](https://arxiv.org/abs/1406.2661) [[stat.ML](https://arxiv.org/abs/1406.2661)].
- [21] Tero Karras, Samuli Laine, and Timo Aila. *A Style-Based Generator Architecture for Generative Adversarial Networks*. 2019. arXiv: [1812.04948](https://arxiv.org/abs/1812.04948) [[cs.NE](https://arxiv.org/abs/1812.04948)].
- [22] Robin Rombach et al. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2022. arXiv: [2112.10752](https://arxiv.org/abs/2112.10752) [[cs.CV](https://arxiv.org/abs/2112.10752)].
- [23] Mehdi Mirza and Simon Osindero. *Conditional Generative Adversarial Nets*. 2014. arXiv: [1411.1784](https://arxiv.org/abs/1411.1784) [[cs.LG](https://arxiv.org/abs/1411.1784)].
- [24] Scott Reed et al. *Generative Adversarial Text to Image Synthesis*. 2016. arXiv: [1605.05396](https://arxiv.org/abs/1605.05396) [[cs.NE](https://arxiv.org/abs/1605.05396)].
- [25] Phillip Isola et al. *Image-to-Image Translation with Conditional Adversarial Networks*. 2018. arXiv: [1611.07004](https://arxiv.org/abs/1611.07004) [[cs.CV](https://arxiv.org/abs/1611.07004)].
- [26] Jascha Sohl-Dickstein et al. *Deep Unsupervised Learning using Nonequilibrium Thermodynamics*. 2015. arXiv: [1503.03585](https://arxiv.org/abs/1503.03585) [[cs.LG](https://arxiv.org/abs/1503.03585)].
- [27] Prafulla Dhariwal and Alex Nichol. *Diffusion Models Beat GANs on Image Synthesis*. 2021. arXiv: [2105.05233](https://arxiv.org/abs/2105.05233) [[cs.LG](https://arxiv.org/abs/2105.05233)].
- [28] Jonathan Ho and Tim Salimans. *Classifier-Free Diffusion Guidance*. 2022. arXiv: [2207.12598](https://arxiv.org/abs/2207.12598) [[cs.LG](https://arxiv.org/abs/2207.12598)].
- [29] Patrick Esser, Robin Rombach, and Björn Ommer. *Taming Transformers for High-Resolution Image Synthesis*. 2021. arXiv: [2012.09841](https://arxiv.org/abs/2012.09841) [[cs.CV](https://arxiv.org/abs/2012.09841)].
- [30] Diederik P Kingma and Max Welling. *Auto-Encoding Variational Bayes*. 2022. arXiv: [1312.6114](https://arxiv.org/abs/1312.6114) [[stat.ML](https://arxiv.org/abs/1312.6114)].

- [31] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. *Neural Discrete Representation Learning*. 2018. arXiv: [1711.00937](https://arxiv.org/abs/1711.00937) [cs.LG].
- [32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: [1505.04597](https://arxiv.org/abs/1505.04597) [cs.CV].
- [33] Thomas Wolf et al. *HuggingFace’s Transformers: State-of-the-art Natural Language Processing*. 2020. arXiv: [1910.03771](https://arxiv.org/abs/1910.03771) [cs.CL].
- [34] Mathias Eitz et al. “Sketch-Based Image Retrieval: Benchmark and Bag-of-Features Descriptors”. In: *IEEE Transactions on Visualization and Computer Graphics* 17.11 (2011), pp. 1624–1636. DOI: [10.1109/TVCG.2010.266](https://doi.org/10.1109/TVCG.2010.266).
- [35] Rui Hu, Mark Barnard, and John Collomosse. “Gradient field descriptor for sketch based retrieval and localization”. In: Sept. 2010, pp. 1025–1028. DOI: [10.1109/ICIP.2010.5649331](https://doi.org/10.1109/ICIP.2010.5649331).
- [36] Yang Cao et al. “Edgel index for large-scale sketch-based image search”. In: *CVPR 2011*. 2011, pp. 761–768. DOI: [10.1109/CVPR.2011.5995460](https://doi.org/10.1109/CVPR.2011.5995460).
- [37] Rui Hu, Tinghuai Wang, and John Collomosse. “A bag-of-regions approach to sketch-based image retrieval”. In: *2011 18th IEEE International Conference on Image Processing*. 2011, pp. 3661–3664. DOI: [10.1109/ICIP.2011.6116513](https://doi.org/10.1109/ICIP.2011.6116513).
- [38] Stuart James, Manuel J. Fonseca, and John Collomosse. “ReEnact: Sketch based Choreographic Design from Archival Dance Footage”. In: Apr. 2014. DOI: [10.1145/2578726.2578766](https://doi.org/10.1145/2578726.2578766).
- [39] Daniyar Turmukhambetov et al. “Interactive Sketch-Driven Image Synthesis”. In: *Computer Graphics Forum* 34 (Aug. 2015). DOI: [10.1111/cgf.12665](https://doi.org/10.1111/cgf.12665).
- [40] Fang Wang, Le Kang, and Yi Li. “Sketch-based 3D shape retrieval using Convolutional Neural Networks”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)*, pp. 1875–1883. URL: [\newlinehttps://api.semanticscholar.org/CorpusID:14343656](https://api.semanticscholar.org/CorpusID:14343656).
- [41] Mehdi Mirza and Simon Osindero. “Conditional Generative Adversarial Nets”. In: *CoRR* abs/1411.1784 (2014). arXiv: [1411.1784](https://arxiv.org/abs/1411.1784). URL: <http://arxiv.org/abs/1411.1784>.
- [42] Chengying Gao et al. “SketchyCOCO: Image Generation From Freehand Scene Sketches”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 5173–5182. DOI: [10.1109/CVPR42600.2020.00522](https://doi.org/10.1109/CVPR42600.2020.00522).
- [43] Arnab Ghosh et al. *Interactive Sketch & Fill: Multiclass Sketch-to-Image Translation*. 2019. arXiv: [1909.11081](https://arxiv.org/abs/1909.11081) [cs.CV].
- [44] Runtao Liu, Qian Yu, and Stella Yu. *Unsupervised Sketch-to-Photo Synthesis*. 2020. arXiv: [1909.08313](https://arxiv.org/abs/1909.08313) [cs.CV].
- [45] Jun-Yan Zhu et al. *Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks*. 2020. arXiv: [1703.10593](https://arxiv.org/abs/1703.10593) [cs.CV].
- [46] Peng Xu et al. “SketchMate: Deep Hashing for Million-Scale Human Sketch Retrieval”. In: *CoRR* abs/1804.01401 (2018). arXiv: [1804.01401](https://arxiv.org/abs/1804.01401). URL: <http://arxiv.org/abs/1804.01401>.

- [47] Shu-Yu Chen et al. “DeepFaceDrawing: deep generation of face images from sketches”. In: *ACM Transactions on Graphics* 39 (July 2020). DOI: [10.1145/3386569.3392386](https://doi.org/10.1145/3386569.3392386).
- [48] Yuhang Li et al. “DeepFacePencil: Creating Face Images from Freehand Sketches”. In: *Proceedings of the 28th ACM International Conference on Multimedia*. MM ’20. ACM, Oct. 2020. DOI: [10.1145/3394171.3413684](https://doi.org/10.1145/3394171.3413684). URL: <http://dx.doi.org/10.1145/3394171.3413684>.
- [49] Luke Metz et al. “Unrolled Generative Adversarial Networks”. In: *CoRR* abs/1611.02163 (2016). arXiv: [1611.02163](https://arxiv.org/abs/1611.02163). URL: <http://arxiv.org/abs/1611.02163>.
- [50] Qiang Wang et al. *DiffSketching: Sketch Control Image Synthesis with Diffusion Models*. May 2023.
- [51] Dmitry Baranchuk et al. *Label-Efficient Semantic Segmentation with Diffusion Models*. 2022. arXiv: [2112.03126](https://arxiv.org/abs/2112.03126) [cs.CV].
- [52] Edgar Simo-Serra et al. “Learning to simplify: fully convolutional networks for rough sketch cleanup”. In: *ACM Transactions on Graphics* 35 (July 2016), pp. 1–11. DOI: [10.1145/2897824.2925972](https://doi.org/10.1145/2897824.2925972).
- [53] Edgar Simo-Serra, Satoshi Iizuka, and Hiroshi Ishikawa. *Mastering Sketching: Adversarial Augmentation for Structured Prediction*. 2017.
- [54] Olga Russakovsky et al. *ImageNet Large Scale Visual Recognition Challenge*. 2015. arXiv: [1409.0575](https://arxiv.org/abs/1409.0575) [cs.CV].
- [55] Zhuo Su et al. *Pixel Difference Networks for Efficient Edge Detection*. 2021. arXiv: [2108.07009](https://arxiv.org/abs/2108.07009) [cs.CV].
- [56] Patsorn Sangkloy et al. “The sketchy database: learning to retrieve badly drawn bunnies”. In: *ACM Transactions on Graphics* 35 (July 2016), pp. 1–12. DOI: [10.1145/2897824.2925954](https://doi.org/10.1145/2897824.2925954).