National Technical University of Athens
School of Electrical and Computer Engineering
Division of Computer Science
Artificial Intelligence and Learning Systems Lab

Cross-modal Flow Matching for Domain Generalization

Diploma Thesis

of

ANTONIOS KRITIKOS

Supervisor: Athanasios Voulodimos
Assistant Professor, NTUA

Athens, November 2025

National Technical University of Athens
School of Electrical and Computer Engineering
Division of Computer Science
Artificial Intelligence and Learning Systems Lab

Cross-modal Flow Matching for Domain Generalization

Diploma Thesis

of

ANTONIOS KRITIKOS

Supervisor: Athanasios Voulodimos
Assistant Professor, NTUA

Approved by the examination committee on 11th November 2025.

(Signature)	*(Signature)*	*(Signature)*
.....	.....	.....
Athanasios Voulodimos	Giorgos Stamou	Andreas-Georgios Stafylopatis
Assistant Professor, NTUA	Professor, NTUA	Professor Emeritus, NTUA

Athens, November 2025

National Technical University of Athens
School of Electrical and Computer Engineering
Division of Computer Science
Artificial Intelligence and Learning Systems Lab

DISCLAIMER ON ACADEMIC ETHICS AND INTELLECTUAL PROPERTY RIGHTS

Being fully aware of the implications of copyright laws, I expressly state that this diploma thesis, as well as the electronic files and source codes developed or modified in the course of this thesis, are solely the product of my personal work and do not infringe any rights of intellectual property, personality and personal data of third parties, do not contain work / contributions of third parties for which the permission of the authors / beneficiaries is required and are not a product of partial or complete plagiarism, while the sources used are limited to the bibliographic references only and meet the rules of scientific citing. The points where I have used ideas, text, files and / or sources of other authors are clearly mentioned in the text with the appropriate citation and the relevant complete reference is included in the bibliographic references section. I fully, individually and personally undertake all legal and administrative consequences that may arise in the event that it is proven, in the course of time, that this thesis or part of it does not belong to me because it is a product of plagiarism.

(Signature)

.....
Antonios Kritikos

Graduate of School of Electrical and Computer Engineering, National Technical University of Athens

11th November 2025

Περίληψη

Η γενίκευση τομέων (ΓΤ) επιτάσσει εύρωστη λειτουργία των μοντέλων υπό συνθήκες μετατόπισης τομέα, η οποία στο πεδίο της όρασης υπολογιστών εκδηλώνεται κυρίως ως στιλιστικές παραλλαγές μεταξύ δειγμάτων. Παρά την εντατική έρευνα, τα μοντέλα συχνά υπερπροσαρμόζονται σε χαρακτηριστικά εμφάνισης συγκεκριμένου τομέα και δεν καταφέρνουν να αποτυπώσουν τη σημασιολογία της κατηγορίας. Ως εκ τούτου, έχουν καταβληθεί πολλές προσπάθειες για την ενσωμάτωση της φυσικής γλώσσας, λόγω της εγγενούς αμεταβλητότητάς της σε σχέση με τον τομέα. Ωστόσο, οι τρέχουσες πολυτροπικές προσεγγίσεις, και συγκεκριμένα εκείνες που βασίζονται σε ομοιότητα συνημιτόνου για την αντιπαραθετική ευθυγράμμιση σε έναν κοινό λανθάνοντα χώρο, υποφέρουν από το «τροπικό κενό», ένα φαινόμενο στο οποίο οι αναπαραστάσεις εικόνας και κειμένου καταλαμβάνουν διακριτές περιοχές παρά τη σημασιολογική ευθυγράμμιση. Σε αυτή την εργασία, αντιμετωπίζουμε αυτό το κενό εφαρμόζοντας αντιστοίχιση ροής για να μάθουμε έναν συνεχή μετασχηματισμό στον από κοινού Ευκλείδειο διανυσματικό χώρο μεταξύ μη κανονικοποιημένων αναπαραστάσεων εικόνας και κειμένου της ίδιας κατηγορίας. Αποκλίνοντας από τη συνήθη πρακτική των απλών κατανομών πηγής, εκπαιδεύουμε ένα διανυσματικό πεδίο που προκαλεί τη ροή αναπαραστάσεων εικόνων με πιθανή επίδραση του τομέα προς αναπαραστάσεις κειμένου ανεξάρτητες από τον τομέα. Το προκύπτον πλαίσιο, ονόματι *CrossFlowDG*, αξιολογείται χρησιμοποιώντας τον αποδοτικό κωδικοποιητή εικόνων VMamba και σημειώνει κορυφαία ακρίβεια κατηγοριοποίησης σε διάφορους απαιτητικούς τομείς. Για να ενισχύσουμε περαιτέρω την αποδοτικότητα του συμπερασμού και, ως εκ τούτου, την δυνατότητα διάθεσης σε συσκευές με περιορισμένες υπολογιστικές δυνατότητες, προτείνουμε μια παραλλαγή βασισμένη στη βέλτιστη μεταφορά, που ονομάζεται *OT-CrossFlowDG*. Το δεύτερο αυτό πλαίσιο ενσωματώνει έναν μηχανισμό ευθυγράμμισης βέλτιστης μεταφοράς που ελαχιστοποιεί την απόσταση 2-Wasserstein μεταξύ των κατανομών εικόνων και κειμένων κάθε κατηγορίας στον Ευκλείδειο χώρο. Επιπλέον, βάσει του αλγορίθμου Sinkhorn-Knopp δίνει μια κατά προσέγγιση βέλτιστη σύζευξη μεταξύ δειγμάτων εικόνων και κειμένων της ίδιας κατηγορίας και χρησιμοποιεί τους προκύπτοντες βαρυκεντρικούς στόχους ως βελτιωμένα σημεία επίβλεψης της ροής. Ως αποτέλεσμα, το *OT-CrossFlowDG* επιτυγχάνει τη μέγιστη επίδοσή του χρησιμοποιώντας μόνο 1 βήμα συμπερασμού, υπογραμμίζοντας την αποδοτικότητά του σε σενάρια με υπολογιστικούς περιορισμούς.

Λέξεις Κλειδιά

μηχανική μάθηση, όραση υπολογιστών, γενίκευση τομέων, αντιπαραθετική μάθηση, τροπικό κενό, αντιστοίχιση ροής, βέλτιστη μεταφορά

Abstract

Domain generalization (DG) requires models to perform robustly under domain shift, which in the realm of computer vision mainly takes form as stylistic variations between samples. Despite intensive research, DG still poses a major challenge, as models often overfit to domain-specific appearance cues and fail to capture class semantics. Therefore, many efforts have explored the use of natural language, due to its inherent domain-invariance. However, current multimodal approaches, specifically those relying on cosine similarity for cross-modal contrastive alignment in a joint embedding space, suffer from the *modality gap*, a phenomenon where image and text embeddings occupy separate regions despite semantic alignment. In this thesis, we address this residual gap by applying flow matching to learn a continuous transformation between unnormalized image and text embeddings of the same class, in the joint Euclidean latent space. Unlike most prior work which uses simple source distributions, we instead train a vector field that explicitly flows potentially domain-biased image embeddings to domain-invariant text embeddings. The resulting framework, *CrossFlowDG*, is tested with the efficient VMamba image encoder, which achieves linear complexity, compared to the widely-used quadratic-complexity transformer backbones, and establishes state-of-the-art classification accuracy among similar methods across several challenging domains from relevant benchmarks. To further enhance inference efficiency, and therefore deployability on edge devices with restricted compute, we propose an optimal transport-informed variant, called *OT-CrossFlowDG*. This second framework incorporates optimal transport alignment to minimize the 2-Wasserstein distance between modality distributions of each class in the Euclidean space. By solving the Sinkhorn-Knopp problem to obtain approximate optimal couplings between image and text embeddings of the same class and using barycentric targets as refined flow supervision, *OT-CrossFlowDG* achieves its peak performance in only 1 inference step, highlighting its efficiency in computationally constrained deployment scenarios.

Keywords

machine learning, computer vision, domain generalization, contrastive learning, modality gap, flow matching, optimal transport

στην οικογένεια και
τους φίλους μου

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω από καρδιάς τον επιβλέποντα καθηγητή κ. Θάνο Βουλόδημο για την καθοδήγηση στο πλαίσιο αυτής της εργασίας, αλλά και για τις πολύτιμες πάσης φύσεως συμβουλές. Επίσης οφείλω ένα πολύ μεγάλο «ευχαριστώ» στον υποψήφιο διδάκτορα Νίκο Σπανό για τα brainstorming sessions μας και τη στενή συνεργασία καθ' όλη τη διάρκεια των τελευταίων μηνών.

Τίποτα, ωστόσο, δε θα είχε επιτευχθεί χωρίς τη συμβολή και την ενθάρρυνση των γονιών μου, Γιάννη και Σοφίας, και της αδερφής μου, Καλλιόπης. Τέλος, νιώθω ευγνώμων για όσες φιλίες –τόσο εντός όσο κι εκτός σχολής– έχουν χτιστεί και γκρεμιστεί όλα αυτά τα χρόνια, καθώς έχουν συμβάλει καθοριστικά στη διαμόρφωση του σημερινού μου εαυτού.

Η εργασία αυτή δε θεωρώ ότι κλείνει κάποιον κύκλο· αντιθέτως, ολοκληρώνει μια σπείρα στην εξέλιξή μου ως ερευνητή και ως ανθρώπου. Δεν μπορώ παρά να αισθάνομαι μια άγρια ανυπομονησία για το μέλλον, που στέκεται μπροστά μας αδιάφανο και ανεξερεύνητο...

AWS resources were provided by the National Infrastructures for Research and Technology GRNET and funded by the EU Recovery and Resiliency Facility

Table of Contents

List of Figures

List of Tables

Chapter 1

Εκτεταμένη περίληψη στα ελληνικά

1.1 Εισαγωγή

Η Γενίκευση Τομέων (ΓΤ) αφορά στην ικανότητα των μοντέλων να διατηρούν υψηλή επίδοση υπό συνθήκες μετατόπισης τομέα, η οποία στην όραση υπολογιστών εκδηλώνεται κυρίως ως στιλιστικές παραλλαγές μεταξύ δειγμάτων, χωρίς πρόσβαση σε δεδομένα από τον τομέα στόχου κατά την εκπαίδευση. Η ικανότητα αυτή της γενίκευσης είναι ιδιαίτερα κρίσιμη διότι απαντάται διαρκώς σε πρακτικές εφαρμογές, όπως αυτόνομη οδήγηση και ιατρικά διαγνωστικά συστήματα, όπου το εκάστοτε μοντέλο καλείται να αποδώσει σε περιβάλλοντα ενδεχομένως πολύ πιο δυσχερή από εκείνα στα οποία έχει εκπαιδευτεί. Αν και έχουν αναπτυχθεί διάφορες προσεγγίσεις για τη ΓΤ, όπως επαύξηση δεδομένων και μετα-μάθηση, ιδιαίτερο ενδιαφέρον παρουσιάζουν οι πολυτροπικές μέθοδοι, διότι συνδυάζουν όραση, γλώσσα ή και άλλες τροπικότητες για την εκμάθηση πιο εύρωστων αναπαραστάσεων υπό συνθήκες μετατόπισης τομέα. Όμως, οι πολυτροπικές προσεγγίσεις, και συγκεκριμένα εκείνες που στηρίζονται στην ευθυγράμμιση αναπαραστάσεων σε έναν κοινό διανυσματικό χώρο, παρουσιάζουν το φαινόμενο του «τροπικού κενού»: οι ενσωματώσεις εικόνας και κειμένου καταλήγουν σε ξεχωριστές περιοχές του χώρου, εμποδίζοντας συχνά την αξιοποίηση του αναλλοίωτου της κειμενικής πληροφορίας. Στο πλαίσιο της ΓΤ, η ύπαρξη του τροπικού κενού εγείρει ένα βασικό ερώτημα:

Μπορεί η γεφύρωση του τροπικού κενού να ενισχύσει
την ικανότητα γενίκευσης του μοντέλου;

Αυτή η εργασία διερευνά αν η Αντιστοίχιση Ροής (ΑΡ), που στοχεύει στην εκμάθηση συνεχών μετασχηματισμών μεταξύ κατανομών, μπορεί να γεφυρώσει το τροπικό κενό και να ενισχύσει την ικανότητα ΓΤ, συνδέοντας τις αναπαραστάσεις της εικόνας με αυτές του κειμένου.

Οι βασικότερες συνεισφορές αυτής της εργασίας συνοψίζονται ως εξής:

1. Θίγουμε την πρόκληση της ευθυγράμμισης των αναπαραστάσεων εικόνας-κειμένου εισάγοντας ΑΡ χωρίς θόρυβο για τη γεφύρωση του τροπικού κενού.

2. Προτείνουμε το *CrossFlowDG*, ένα καινοτόμο πλαίσιο ΓΤ που συνδυάζει την αντιπαραθετική μάθηση ομοιότητας συνημιτόνου για μια πρωτογενή διατροπική ευθυγράμμιση με την ΑΡ για να γεφυρώσει το εναπομείναν κενό. Επιπλέον, το πλαίσιό μας δεν

υποθέτει ευθυγράμμιση των τροπικοτήτων κατά την προεκπαίδευση, σε αντίθεση με μεγάλο κομμάτι της σχετικής βιβλιογραφίας. Χρησιμοποιώντας το προεκπαιδευμένο VMamba, το *CrossFlowDG* επιτυγχάνει κορυφαία επίδοση σε πολλούς τομείς επιδεικνύοντας αποδοτικότητα κατά τον συμπερασμό.

3. Επίσης, εισάγουμε το *OT-CrossFlowDG*, μια παραλλαγή του *FlowDG* που ενσωματώνει έναν αλγόριθμο βέλτιστης μεταφοράς ανά κατηγορία για να εξαγάγει προσεγγιστικά βέλτιστα ζεύγη μεταξύ αναπαραστάσεων εικόνων και βαρυκεντρικών αναπαραστάσεων κειμένου, που δρουν ως βελτιωμένα σημεία επίβλεψης για την καθοδήγηση της διαδικασίας AP. Ως αποτέλεσμα, το *OT-CrossFlowDG* επιτυγχάνει τη μέγιστη επίδοσή του σε 1 μόνο βήμα συμπερασμού, καθιστώντας το πιο κατάλληλο για διάθεση σε σύνθηκες με υπολογιστικούς περιορισμούς.

1.2 Θεωρητικό υπόβαθρο

1.2.1 Γενίκευση τομέων

Διατύπωση προβλήματος

Στο πρόβλημα της Γενίκευσης Τομέων (ΓΤ, αγγλ. «domain generalization»), θεωρούμε ότι τα δεδομένα εκπαίδευσης (training) και δοκιμής (test) προέρχονται από διαφορετικές αλλά συναφείς κατανομές (τομείς) [1, 13]. Οπτικά, αυτό αποτυπώνεται με στιλιστικές διαφορές μεταξύ των εικόνων διαφορετικών τομέων (π.χ., οι διαφορές μεταξύ μιας φωτογραφίας και ενός σκίτσου μιας γάτας, βλ. επίσης το Σχήμα 1.1).

Figure 1.1. *Παραδείγματα μετατόπισης τομέα σε δύο συνήθη σύνολα αξιολόγησης της γενίκευσης τομέων. Πηγή: [1]*

Κατά την εκπαίδευση, έχουμε πρόσβαση σε επισημειωμένα δεδομένα από πολυάριθμους τομείς εκπαίδευσης (source domains):

$$\mathcal{S} = \bigcup_{i=1}^{N} \mathcal{S}_i = \bigcup_{i=1}^{N} \{(x_j^{(i)}, y_j^{(i)})\}_{j=1}^{n_i}, \tag{1.1}$$

όπου το σύνολο $\mathcal{S}_i = \{(x_j^{(i)}, y_j^{(i)})\}_{j=1}^{n_i}$ σημαίνει το σύνολο εκπαίδευσης του τομέα $\mathcal{D}_i$ με n_i δείγματα και $(x_j^{(i)}, y_j^{(i)}) \sim P_i(X, Y)$.

Στόχος της ΓΤ είναι η εκμάθηση μιας συνάρτησης πρόβλεψης $f : \mathcal{X} \to \mathcal{Y}$, η οποία γενικεύει ικανοποιητικά σε άγνωστο τομέα στόχου (target domain) $\mathcal{D}_t$ με κατανομή $P_t(X, Y)$, όπου $\mathcal{D}_t \notin \mathcal{D}$. Τυπικά, στοχεύουμε στην ελαχιστοποίηση του προσδοκώμενου ρίσκου επάνω στον τομέα στόχου:

$$R_t(f) = \mathbb{E}_{(x,y)\sim P_t(X,Y)}[\ell(f(x), y)], \tag{1.2}$$

όπου η $\ell(\cdot, \cdot)$ είναι μια συνάρτηση κόστους.

Μιας και **δε διαθέτουμε πρόσβαση στα δεδομένα του τομέα στόχου κατά την εκπαίδευση**, ελαχιστοποιούμε το μέσο εμπειρικό ρίσκο επάνω σε όλους τους διαθέσιμους τομείς εκπαίδευσης:

$$\hat{R}_{\text{emp}}(f) = \frac{1}{|\mathcal{S}|} \sum_{i=1}^{N} \sum_{j=1}^{n_i} \ell(f(x_j^{(i)}), y_j^{(i)}). \tag{1.3}$$

Εντούτοις, η ελαχιστοποίηση του εμπειρικού ρίσκου ενδέχεται να οδηγήσει σε υπερπροσαρμογή στους τομείς εκπαίδευσης. Γι' αυτό, οι μέθοδοι ΓΤ συνήθως βελτιστοποιούν μια ομαλοποιημένη εκδοχή:

$$\min_f \hat{R}_{\text{emp}}(f) + \lambda \Omega(f), \tag{1.4}$$

όπου ο $\Omega(f)$ είναι ένας όρος ομαλοποίησης που ενθαρρύνει αναπαραστάσεις οι οποίες παραμένουν αναλλοίωτες συναρτήσει του τομέα και το $\lambda > 0$ είναι μια παράμετρος που ελέγχει τον βαθμό της ομαλοποίησης.

Το βασικότερο χαρακτηριστικό στις κλασικές εφαρμογές ΓΤ είναι η έλλειψη πρόσβασης στα δεδομένα τομέα στόχου κατά την εκπαίδευση. Επιπλέον, στην πλειονότητα των εφαρμογών υποτίθεται ότι ο χώρος επισημειώσεων του τομέα δοκιμής παραμένει ίδιος με αυτόν των τομέων εκπαίδευσης. Την παραπάνω υπόθεση υιοθετούμε και στη μεθοδολογία που αναπτύσσουμε στο κεφάλαιο 4.

Ωστόσο, τυγχάνει να διανύουμε την εποχή της μαζικής προεκπαίδευσης (pretraining), δηλαδή της εκπαίδευσης μοντέλων γενικής χρήσης με ογκώδη σύνολα δεδομένων. Τα σύνολα αυτά είναι κυρίως εξορυγμένα από το διαδίκτυο, χωρίς ιδιαίτερη διήθηση των διπλοτύπων ή δειγμάτων που ανήκουν σε συγκεκριμένο τομέα. Από αυτή την οπτική, τέτοια σύνολα εκπαίδευσης θεωρούνται απο κάποιους «μολυσμένα» και άρα ακατάλληλα για εφαρμογές ΓΤ, αφού δεν υπάρχουν εγγυήσεις για το «άβατο» του τομέα στόχου. Για τον λόγο αυτό, η επιστημονική κοινότητα στρέφεται προς την κατασκευή πιο καθαρών συνόλων προεκπαίδευσης και πιο απαιτητικών συνόλων αξιολόγησης [14].

Μεθοδολογίες

Οι προσεγγίσεις ΓΤ μπορούν, βάσει του ειδικού τους στόχου, να ταξινομηθούν σε πληθώρα κατηγοριών [1], όπως είναι η **ευθυγράμμιση τομέων** (domain alignment), μέσω ανταγωνιστικής (adversarial) εκπαίδευσης [15, 16, 17] ή ελαχιστοποίησης κάποιας μετρικής απόστασης [18, 19], η **επαύξηση δεδομένων** στον χώρο των εικόνων [20, 21] ή των χαρακτηριστικών [22, 23], οι εναλλακτικές **στρατηγικές εκμάθησης** [24, 25, 26, 27, 28, 29, 30], η **εκμάθηση** αιτιολογικών και αποσυζευγμένων **αναπαραστάσεων** [31, 32, 33], καθώς και οι **αρχιτεκτονικές μέθοδοι** [34, 35, 36, 37].

Πρωτόκολλο αξιολόγησης

Όπως είθισται στις περισσότερες εφαρμογές ΓΤ, υιοθετούμε το «leave-one-domain-out» πρωτόκολλο αξιολόγησης, δηλαδή, αν το σύνολο εκπαίδευσης περιλαμβάνει N τομείς, εκπαιδεύουμε το μοντέλο σε $N-1$ τομείς εκπαίδευσης και δοκιμάζουμε το μοντέλο στον εναπομείναντα τομέα:

$$\text{LODO}(f) = \frac{1}{N}\sum_{i=1}^{N} R_i(f_{\setminus i}), \tag{1.5}$$

όπου το μοντέλο $f_{\setminus i}$ έχει εκπαιδευτεί στο σύνολο $\mathcal{D} \setminus \{\mathcal{D}_i\}$.

1.2.2 Μοντέλα χώρου καταστάσεων

Το πεδίο της Μηχανικής Μάθησης (ΜΜ) γνώρισε μεγάλη άνθηση χάρη στην αρχιτεκτονική του μετασχηματιστή (transformer) μέσω του μηχανισμού προσοχής (attention), ο οποίος λαμβάνει υπόψη την αλληλεπίδραση κάθε ζεύγους στοιχείων μιας ακολουθίας [38]. Ωστόσο, η **τετραγωνική πολυπλοκότητα** $O(L^2)$ του **μετασχηματιστή** συναρτήσει του μήκους L της ακολουθίας εισόδου ώθησε την ερευνητική κοινότητα στην εξεύρεση αποδοτικότερων εναλλακτικών.

Τα **Μοντέλα Χώρου Καταστάσεων** (ΜΧΚ) εκ πρώτης όψεως προσφέρουν **γραμμική πολυπλοκότητα** $O(L)$ επεξεργασίας της ακολουθίας εισόδου μήκους L. Εντούτοις, στην κλασική τους έκφανση προβάλλουν μια σειρά ασυμβατοτήτων που σχετίζονται κυρίως με την παράλληλη φύση των σύγχρονων αρχιτεκτονικών Μονάδων Επεξεργασίας Γραφικών (ΜΕΓ, αγγλ. «Graphics Processing Unit», «GPU»).

Η αρχιτεκτονική Mamba αποτελεί σημείο καμπής στην υιοθέτηση των ΜΧΚ σε εφαρμογές ΜΜ, καθώς προτείνει ένα ΜΧΚ που μπορεί **επιλεκτικά** να συγκρατεί ή να απορρίπτει πρότερη πληροφορία και να **εκπαιδεύεται αποδοτικά** σε σύγχρονες ΜΕΓ [2].

Μαθηματική διατύπωση

Έχοντας τις καταβολές τους στην κλασική θεωρία ελέγχου, τα ΜΧΚ είναι μια κατηγορία μοντέλων που αντιστοιχίζουν ένα μονοδιάστατο (1Δ) σήμα εισόδου $u(t)$ σε ένα επίσης 1Δ σήμα εξόδου $y(t)$ μέσω μιας ενδιάμεσης πολυδιάστατης λανθάνουσας κατάστασης $x(t) \in \mathbb{R}^N$.

Στη διατύπωση **συνεχούς χρόνου**, το σύστημα λαμβάνει τη μορφή μιας πρωτοβάθμιας Συνήθους Διαφορικής Εξίσωσης (ΣΔΕ):

$$\frac{dx(t)}{dt} = \mathbf{A}x(t) + \mathbf{B}u(t) \tag{1.6}$$

$$y(t) = \mathbf{C}x(t) + \mathbf{D}u(t) \tag{1.7}$$

Ωστόσο, μιας και οι περισσότερες πρακτικές εφαρμογές ΜΜ αφορούν σε συστήματα **διακριτού χρόνου**, μετατρέπουμε αναλόγως το σύστημα 1.6-1.7, μέσω κάποιας μεθόδου διακριτοποίησης, π.χ., συγκράτηση μηδενικής τάξης (Zero-Order Hold, ZOH):

$$x_k = \bar{\mathbf{A}}x_{k-1} + \bar{\mathbf{B}}u_k \tag{1.8}$$

$$y_k = \mathbf{C}x_k + \mathbf{D}u_k \tag{1.9}$$

Οι διακριτοποιημένοι πίνακες $\bar{\mathbf{A}}$ και $\bar{\mathbf{B}}$ κατασκευάζονται ως εξής:

$$\bar{\mathbf{A}} = e^{\Delta\mathbf{A}}, \qquad \bar{\mathbf{B}} = (\Delta\mathbf{A})^{-1}(e^{\Delta\mathbf{A}} - \mathbf{I})\Delta\mathbf{B}. \tag{1.10}$$

όπου το Δ καλείται βήμα δειγματοληψίας. Η παραπάνω **αναδρομική** μορφή προσφέρεται για αποδοτικό συμπερασμό (inference), λόγω της γραμμικής πολυπλοκότητας συναρτήσει του μήκους εισόδου. Ωστόσο, η μορφή που συνάδει με τις σύγχρονες αρχιτεκτονικές ΜΕΓ είναι η **συνελικτική**:

$$y = \bar{\mathbf{K}} * u, \tag{1.11}$$

όπου ο συνελικτικός πυρήνας $\bar{\mathbf{K}} \in \mathbb{R}^L := (\mathbf{C}\bar{\mathbf{B}}, \mathbf{C}\bar{\mathbf{A}}\bar{\mathbf{B}}, \ldots, \mathbf{C}\bar{\mathbf{A}}^{L-1}\bar{\mathbf{B}})$ εξάγεται από τις παραμέτρους του ΜΧΚ και, εφόσον θεωρείται γνωστός, η έξοδος y μπορεί να υπολογιστεί παράλληλα [7].

Η πορεία προς την αρχιτεκτονική Mamba

Τα **Δομημένα ΜΧΚ** (ΔΜΧΚ, γνωστά ως «S4») [7] υιοθέτησαν (1) την αρχικοποίηση **Προβολικού Τελεστή Υψηλόβαθμου Πολυωνύμου** (αγγλ. «High-Order Polynomial Projection Operator», «HiPPO»), για να διατηρούν μια συνεπτυγμένη αναπαράσταση του ιστορικού μιας συνάρτησης προβάλλοντάς το σε μια βάση ορθογώνιων πολυωνύμων [6], καθώς και (2) τη μορφή **Διαγώνιου Συν Χαμηλού Βαθμού** (πίνακα) για τον πίνακα $\mathbf{A}$, με σκοπό την επιτάχυνση των υπολογισμών που σχετίζονται με αυτόν.

Ωστόσο, οι παράμετροι του ΔΜΧΚ εξακολουθούσαν να παραμένουν χρονικά αναλλοίωτες, περιορίζοντας τις δυνατότητες του μοντέλου. Για τον λόγο αυτό, τα **Επιλεκτικά ΜΧΚ** (ΕΜΧΚ, γνωστά ως «S6») [2] (1) εισήγαγαν έναν **μηχανισμό επιλογής**, με τη δυνατότητα να διατηρεί ή να απορρίπτει πληροφορία βάσει της εισόδου σε κάθε βήμα, και (2) υιοθέτησαν έναν **αλγόριθμο παράλληλης σάρωσης**, με στόχο την αύξηση της αποδοτικότητας.

Εντούτοις, η μετατροπή των ΜΧΚ σε χρονομεταβλητά δημιουργεί μια βασική πρόκληση: η έξοδος δεν μπορεί να υπολογιστεί με μια καθολική συνέλιξη, καθιστώντας την παράλληλη εκπαίδευση δύσχρηστη. Η αρχιτεκτονική Mamba ενσωματώνει το ΕΜΧΚ και αντιμετωπίζει το παραπάνω πρόβλημα με έναν **αλγόριθμο παράλληλης σάρωσης με επίγνωση του υλικού** [2]. Η επιτυχία του αλγορίθμου στηρίζεται σε τρεις τεχνικές:

1. Ο αλγόριθμος πραγματοποιεί **σύντηξη πυρήνων** (υπολογισμού) προκειμένου να μειώσει το πλήθος των εισόδων/εξόδων στη μνήμη, επιτυγχάνοντας σημαντική επιτάχυνση.

2. Οι παράμετροι του ΜΧΚ φορτώνονται κατευθείαν από την αργή Μνήμη Υψηλού Εύρους Ζώνης (ΜΥΕΖ, αγγλ. «High-Bandwidth Memory», «HBM») στη **γρήγορη Στατική Μνήμη Τυχαίας Προσπέλασης** (ΣΜΤΠ, αγγλ. «Static Random-Access Memory», «SRAM»), όπου πραγματοποιούνται η διακριτοποίηση και η αναδρομή. Έπειτα, οι τελικές έξοδοι εγγράφονται στη ΜΥΕΖ.

3. Οι **ενδιάμεσες καταστάσεις** δεν αποθηκεύονται αλλά **επανυπολογίζονται** κατά το οπίσθιο πέρασμα όταν οι είσοδοι φορτώνονται από τη ΜΥΕΖ στη ΣΜΤΠ.

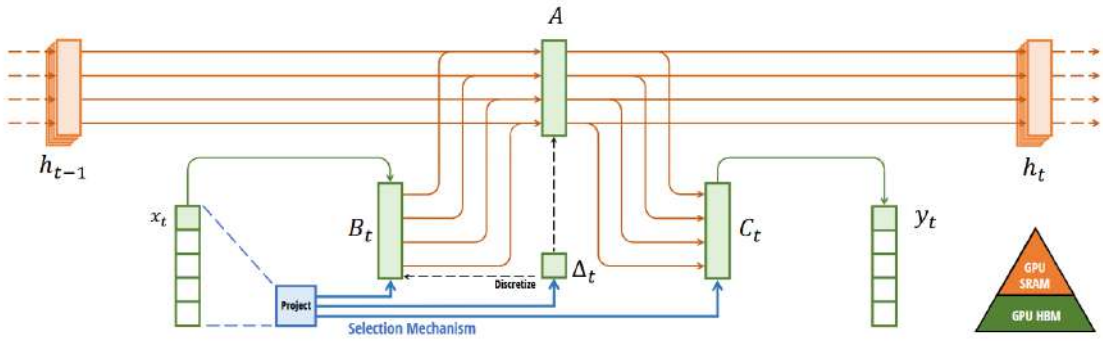


Figure 1.2. *Επισκόπηση της αρχιτεκτονικής Mamba. Ο μηχανισμός επιλογής της αρχιτεκτονικής Mamba συνδυάζει δυναμική του συστήματος που εξαρτάται από την είσοδο με έναν προσεκτικά σχεδιασμένο αλγόριθμο με επίγνωση του υλικού, στοχεύοντας στην υλοποίηση των επεκτεταμένων καταστάσεων αποκλειστικά σε πιο αποδοτικά επίπεδα της ιεραρχίας μνήμης της ΜΕΓ, επιτυγχάνοντας σημαντική επιτάχυνση. Πηγή: [2]*

Η αρχιτεκτονική Mamba στην όραση υπολογιστών

Όπως φανερώνει η πρότερη ανάλυσή μας, τα ΜΧΚ είναι ιδανικά για την επεξεργασία μεγάλων 1Δ ακολουθιών, όπως είναι ένα γραπτό κείμενο. Ωστόσο, η υιοθέτηση των ΜΧΚ σε εφαρμογές Όρασης Υπολογιστών (ΟΥ) δεν είναι τετριμμένη, καθώς αυτές χρησιμοποιούν 2Δ δεδομένα (εικόνες), 3Δ δεδομένα (βίντεο) ή ακόμη πιο σύνθετα δεδομένα υψηλότερης διάστασης. Στη 2Δ περίπτωση, η αφελής προσέγγιση περιλαμβάνει διαχωρισμό της εικόνας σε 2Δ επιθέματα (patches) και ισοπέδωσή τους σε μία διάσταση, μια προσέγγιση που καταργεί την ιδιότητα της 2Δ χωρικής τοπικότητας.

Μια από τις πρώτες επιτυχημένες αρχιτεκτονικές Mamba για εφαρμογές ΟΥ είναι η **Vim** [8], η οποία επεξεργάζεται την ισοπεδωμένη ακολουθία τόσο στην εμπρόσθια όσο και την οπίσθια κατεύθυνση, με στόχο τη διατήρηση των διάφορων χωρικών εξαρτήσεων.

Η αρχιτεκτονική **VMamba** [9] επεκτείνει τη λογική της πολυκατευθυντικής σάρωσης, εφαρμόζοντας ένα σχήμα σάρωσης της εικόνας εισόδου (1) από επάνω αριστερά προς κάτω δεξιά, (2) από κάτω δεξιά προς επάνω αριστερά, (3) από επάνω δεξιά προς κάτω αριστερά και (4) από κάτω αριστερά προς επάνω δεξιά.

Στις εφαρμογές **πυκνής πρόβλεψης**, που απαιτούν κατηγοριοποίηση κάθε ενός εικονοστοιχείου (pixel) της εισόδου, όπως είναι η κατάτμηση εικόνας, συχνά χρησιμοποιούνται αρχιτεκτονικές τύπου U-Net για την ανάκτηση του σχήματος της εικόνας εισόδου στην έξοδο [39, 40, 41]. Άλλες αρχιτεκτονικές είναι **υβριδικές**, δηλαδή συνδυάζουν την αρχιτεκτονική Mamba με Συνελικτικά Νευρωνικά Δίκτυα (ΣΝΔ) [42] ή μετασχηματιστές [43], προκειμένου να δρέψουν τους καρπούς και των δύο μεθόδων.

1.2.3 Αντιπαραθετική μάθηση

Η Αντιπαραθετική Μάθηση (ΑΜ, αγγλ. «Contrastive Learning», «CL») στοχεύει στην εκμάθηση αναπαραστάσεων σε έναν διανυσματικό χώρο, στον οποίο παρόμοια δείγματα απεικονίζονται σε γειτονικά σημεία, ενώ ανόμοια δείγματα σε απομακρυσμένα.

Μαθηματική διατύπωση

Η ΑΜ δρα επάνω σε ζεύγη σημείων δεδομένων. Δεδομένου ενός συνόλου $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$, ορίζουμε τα θετικά ζεύγη (x_i, x_j^+) ως παρόμοια στιγμιότυπα και τα αρνητικά ζεύγη (x_i, x_j^-) ως ανόμοια στιγμιότυπα. Ένα δίκτυο κωδικοποιητή $f_\theta : \mathcal{X} \to \mathbb{R}^d$ αντιστοιχίζει την είσοδο σε έναν d-διάστατο διανυσματικό χώρο, συχνά ακολουθούμενο από μια προβολική κεφαλή $g_\phi : \mathbb{R}^d \to \mathbb{R}^{d'}$, όπου $\mathbb{R}^{d'}$ είναι συνήθως ένας χαμηλότερης διάστασης χώρος στον οποίο εφαρμόζεται η αντιπαραθετική απώλεια. Η κεφαλή αυτή τυπικά παράγει κανονικοποιημένα διανύσματα $z = g_\phi(f_\theta(x))$ με $\|z\| = 1$.

Πυρήνας της ΑΜ είναι η απώλεια InfoNCE [44], η οποία ελαχιστοποιεί την αμοιβαία πληροφορία μεταξύ θετικών ζευγών:

$$\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E}_{(x,x^+)} \left[\log \frac{\exp(\text{sim}(z, z^+)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(z, z_j)/\tau)} \right], \tag{1.12}$$

όπου $z = g_\phi(f_\theta(x))$, όπως ορίζεται παραπάνω, $\text{sim}(z_i, z_j) = z_i^\top z_j$ είναι η **ομοιότητα συνημιτόνου**, τ είναι μια παράμετρος θερμοκρασίας που ελέγχει την πυκνότητα της κατανομής και N είναι το συνολικό πλήθος δειγμάτων στην παρτίδα (batch).

Η παράμετρος θερμοκρασίας τ διαδραματίζει εξέχοντα ρόλο στη διαδικασία βελτιστοποίησης. Χαμηλότερες τιμές παράγουν στενότερες κατανομές που «εξορίζουν» τα άκρως αρνητικά δείγματα, ενώ υψηλότερες τιμές δημιουργούν πιο ομαλές κατανομές. Εμπειρικά, οι τιμές της $\tau \in [0.07, 0.5]$ δίνουν ικανοποιητικά αποτελέσματα σε πολυάριθμες εφαρμογές ενδιαφέροντος [45].

Μεθοδολογίες

Οι αρχικές μέθοδοι ΑΜ απαιτούσαν την ύπαρξη πολλών ρητά αρνητικών δειγμάτων, κάτι που επέβαλλε μεγάλα μεγέθη παρτίδας [45] ή περίπλοκους μηχανισμούς για την παράκαμψη του προβλήματος [46]. Επόμενες προσεγγίσεις υπερέβησαν τα παραπάνω εμπόδια χρησιμοποιώντας δύο αλληλεπιδρώντα δίκτυα που εμποδίζουν την κατάρρευση του μοντέλου, δηλαδή την περίπτωση στην οποία η έξοδος παραμένει σταθερή ανεξαρτήτως εισόδου [47, 48].

Ωστόσο, η αξία της ΑΜ αναδείχθηκε κυρίως μέσω της υιοθέτησής της σε διατροπικές (cross-modal) εφαρμογές, με χαρακτηριστικό παράδειγμα το **CLIP** (Contrastive Language-Image Pretraining) [49]. Το CLIP προεκπαιδεύεται σε ένα ογκώδες και ποικιλόμορφο σύνολο δεδομένων, που συγκροτείται απο 400 εκατομμύρια ζεύγη εικόνων-κειμενικών περιγραφών από το διαδίκτυο. Στο μοντέλο αυτό εκπαιδεύονται από κοινού δύο διακριτοί κωδικοποιητές εικόνας και κειμένου, με χρήση ενός συμμετρικού διατροπικού στόχου ΑΜ, βασισμένου στην εξίσωση 1.12. Η κυριότερη ιδιότητα του CLIP είναι η ικανότητά του για **ταξινόμηση χωρίς παραδείγματα** (zero-shot classification), δηλαδή ταξινόμηση σε κατηγορίες στις οποίες το μοντέλο δεν έχει εκπαιδευθεί, υπολογίζοντας τις πλησιέστερες κειμενικές αναπαραστάσεις (υπό την έννοια της ομοιότητας συνημιτόνου).

Τροπικό κενό

Εμπειρική παρατήρηση. Παρά την επιτυχία μεθόδων όπως το CLIP, η διατροπική ΑΜ συνδέεται με ένα συγκεκριμένο φαινόμενο: αντί οι διανυσματικές αναπαραστάσεις να είναι ομοιόμορφα κατανεμημένες επάνω στη μοναδιαία υπερσφαίρα, συμπιέζονται σε μια στενή, κωνοειδή περιοχή [3]. Αυτό συμβαίνει διότι η απώλεια InfoNCE μεγιστοποιεί την ομοιότητα συνημιτόνου, που ενδιαφέρεται μόνο για τη σχετική γωνία μεταξύ διανυσμάτων και όχι για το μέτρο τους ή την απόλυτη θέση τους στον χώρο. Οπτικά, αν απεικονίζαμε τις διανυσματικές αναπαραστάσεις των δύο τροπικοτήτων (εικόνας και κειμένου), δε θα αντικρίζαμε μια ενιαία συστάδα σημείων, αλλά δύο διακριτές συστάδες – μία για κάθε τροπικότητα – διαχωρισμένες με ένα εμφανές διάστημα, που καλείται **τροπικό κενό** [50].

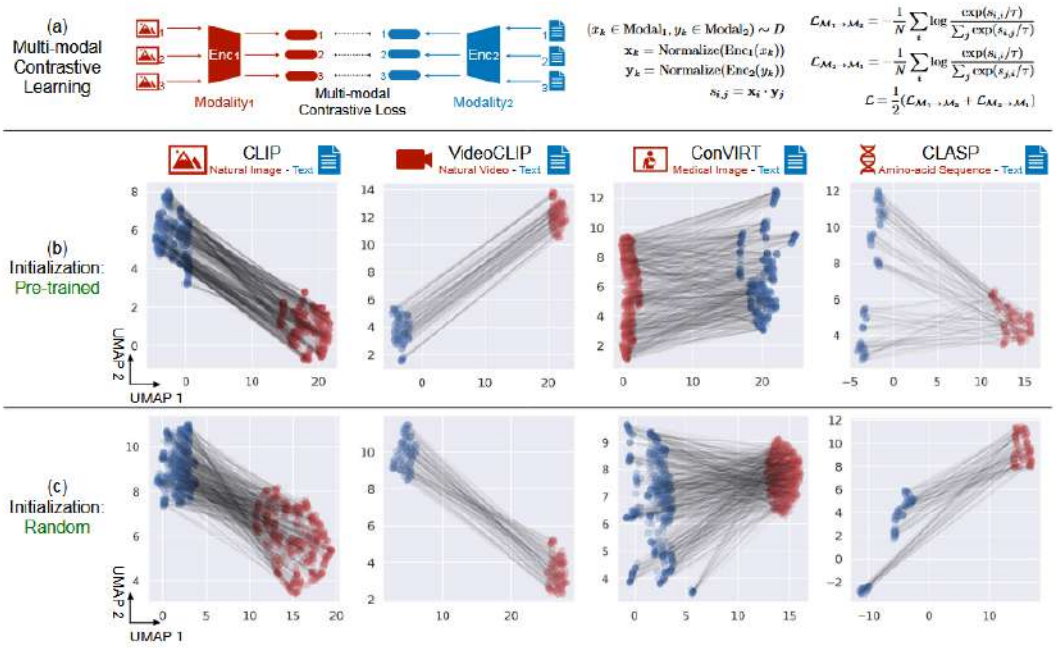


Figure 1.3. *Το τροπικό κενό σε διάφορες εφαρμογές πολυτροπικής αντιπαραθετικής μάθησης. Πηγή: [3]*

Μαθηματική διατύπωση. Δεδομένου ενός συνόλου N ℓ_2-κανονικοποιημένων διανυσματικών αναπαραστάσεων εικόνας $\mathcal{V} = \{\mathbf{v}_i\}_{i=1}^N$ και του αντίστοιχου συνόλου ℓ_2-κανονικοποιημένων διανυσματικών αναπαραστάσεων κειμένου $\mathcal{T} = \{\mathbf{t}_i\}_{i=1}^N$, τα αντίστοιχα κεντροειδή υπολογίζονται ως αλγεβρικοί μέσοι όροι:

$$\boldsymbol{\mu}_{\text{img}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i, \qquad \boldsymbol{\mu}_{\text{txt}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{t}_i. \tag{1.13}$$

Σύμφωνα με το [3], το **τροπικό κενό** μπορεί τυπικά να ποσοτικοποιηθεί ως η ℓ_2-νόρμα της διαφοράς των παραπάνω κεντροειδών:

$$\|\boldsymbol{\Delta}_{gap}\|_2 = \|\boldsymbol{\mu}_{\text{img}} - \boldsymbol{\mu}_{\text{txt}}\|_2. \tag{1.14}$$

Αιτίες και συνέπειες. Η φύση του στόχου ομοιότητας συνημιτόνου, που ενδιαφέρεται μόνο για τη σχετική γωνία μεταξύ διανυσμάτων [45] και οι αρχιτεκτονικές διαφορές μεταξύ των κωδικοποιητών εικόνας και κειμένου [3, 51] αποτελούν τις κυριότερες αιτίες εμφάνισης του τροπικού κενού. Αν και η ύπαρξη του τροπικού κενού μπορεί να περιορίζει ορισμένες δυνατότητες, όπως τη διατροπική παραγωγή [52], εντούτοις ευνοεί τη διατήρηση της λεπτομερούς πληροφορίας που είναι αποκλειστική για κάθε τροπικότητα (π.χ., η περιγραφή «*η φωτογραφία μιας παραλίας που τράβηξα πέρυσι το καλοκαίρι*» περιέχει τη χρονική αναφορά «*πέρυσι το καλοκαίρι*», που δεν είναι οπτικά αντιληπτή) [3]. Άλλες ιδιότητες παραμένουν ανεπηρέαστες, όπως η ικανότητα κατηγοριοποίησης χωρίς παραδείγματα (αγγλ. «zero-shot classification»), διότι βασίζεται στην εύρεση παραπλήσιων αναπαραστάσεων υπό την έννοια της ομοιότητας συνημιτόνου [3].

Γεφυρώνοντας το κενό. Οι σημαντικότερες στρατηγικές γεφύρωσης του τροπικού κενού συμπεριλαμβάνουν εκ των υστέρων κεντροθέτηση των διανυσματικών αναπαραστάσεων καθεμίας τροπικότητας [3], διαμοιρασμό παραμέτρων μεταξύ των κωδικοποιητών εικόνας και κειμένου [51], εκμάθηση συνάρτησης αντιστοίχισης αποκλειστικά για τη διατροπική μεταφορά των αναπαραστάσεων [52] και διάφορα σχήματα προγραμματισμού για τις τιμές της παραμέτρου θερμοκρασίας [53].

1.2.4 Γεννητική μοντελοποίηση

Η γεννητική μοντελοποίηση είναι ένα θεμελιώδες πρότυπο της ΜΜ που στοχεύει στην εκμάθηση της υποκείμενης πραγματικής κατανομής πιθανότητας των παρατηρούμενων δεδομένων, επιτρέποντας τη δημιουργία συνθετικών δειγμάτων που μοιάζουν με την πραγματική κατανομή. Δεδομένου ενός συνόλου δεδομένων $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$ που προέρχεται από μια άγνωστη κατανομή $p_{\text{data}}(x)$, ο στόχος της γεννητικής μοντελοποίησης είναι να προσεγγίσει αυτή την κατανομή με ένα παραμετροποιημένο μοντέλο $p_\theta(x)$ έτσι ώστε τα δείγματα που προέρχονται από την $p_\theta(x)$ να μην μπορούν να διακριθούν από αυτά του αρχικού συνόλου δεδομένων. Αυτό το θεμελιώδες πρόβλημα έχει οδηγήσει σε δεκαετίες έρευνας, από τις πρώτες προσεγγίσεις που βασίζονταν σε Μοντέλα Μίξης Γκαουσιανών (ΜΜΓ, αγγλ. «Gaussian Mixture Models», «GMMs»), τα οποία προσεγγίζουν την πραγματική κατανομή μέσω μιας υπέρθεσης κανονικών κατανομών, και Κρυφά Μαρκοβιανά Μοντέλα (ΚΜΜ, αγγλ. «Hidden Markov Models», «HMMs»), έως τα σύγχρονα πλαίσια βαθιάς μάθησης που μπορούν να δημιουργήσουν εικόνες υψηλής ανάλυσης, ρεαλιστικό κείμενο και σύνθετα δομημένα δεδομένα. Η μαθηματική βάση της γεννητικής μοντελοποίησης βασίζεται στην εκτίμηση πυκνότητας, όπου στόχος είναι να ελαχιστοποιηθεί η απόκλιση μεταξύ της πραγματικής και της μανθανόμενης κατανομής, η οποία συνήθως διατυπώνεται ως μεγιστοποίηση του λογαρίθμου πιθανότητας $\mathcal{L}(\theta) = \mathbb{E}_{x\sim p_{\text{data}}}[\log p_\theta(x)]$ ή, ισοδύναμα, στην ελαχιστοποίηση της απόκλισης Kullback-Leibler $D_{KL}(p_{\text{data}} \| p_\theta) = \mathbb{E}_{x\sim p_{\text{data}}}[\log p_{\text{data}}(x) - \log p_\theta(x)]$ [54, 55].

Στο υπόλοιπο χωρίο, θα περιγράψουμε τα πλέον ευρέως χρησιμοποιούμενα γεννητικά πλαίσια, συμπεριλαμβανομένης της Αντιστοίχισης Ροής (αγγλ. «Flow Matching»), η οποία χρησιμοποιείται στη μέθοδο που προτείνουμε.

Παραλλασσόμενοι Αυτοκωδικοποιητές

Οι **Παραλλασσόμενοι Αυτο-Κωδικοποιητές** (ΠΑΚ, αγγλ. «Variational AutoEncoders», «VAEs») [56, 57, 4] αποτελούν μια συγκροτημένη προσέγγιση της γεννητικής μοντελοποίησης η οποία συνδυάζει τον παραλλασσόμενο συμπερασμό με τη βαθιά μάθηση με στόχο την εκμάθηση μοντέλων λανθανουσών μεταβλητών. Οι ΠΑΚ εισάγουν μια λανθάνουσα μεταβλητή που εγκολπώνει τους υποκείμενους παράγοντες διακύμανσης στα δεδομένα, επιτρέποντας τόσο τον αποδοτικό συμπερασμό όσο και την παραγωγή, μέσω ενός εύχρηστου πλαισίου βελτιστοποίησης.

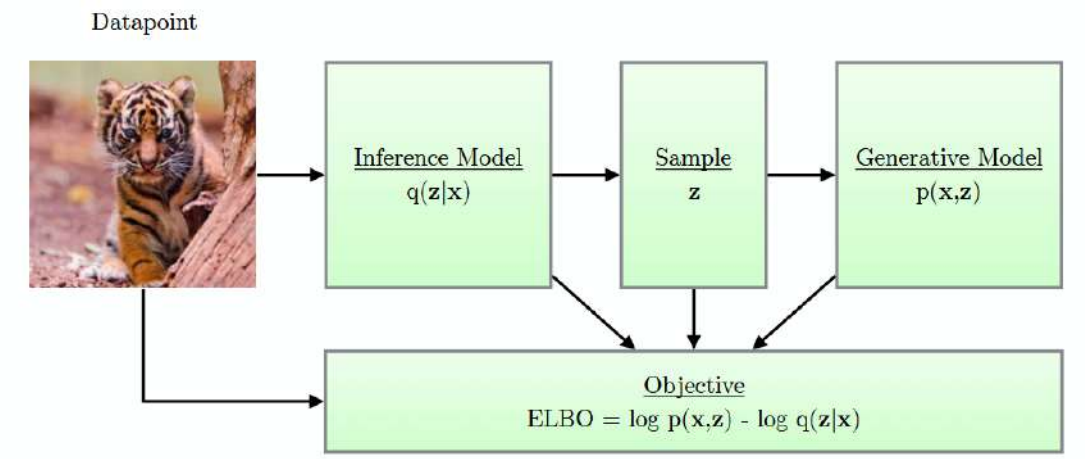


Figure 1.4. *Ένα απλό διάγραμμα υπολογιστικής ροής ενός ΠΑΚ. Πηγή: [4]*

Θεωρητικά θεμέλια. Το πλαίσιο των ΠΑΚ υποθέτει ότι τα παρατηρούμενα δεδομένα x παράγονται από μια λανθάνουσα μεταβλητή z μέσω μιας γεννητικής διαδικασίας $p_\theta(x|z)$, όπου οι λανθάνουσες μεταβλητές ακολουθούν μια εκ των προτέρων κατανομή $p(z)$, που συνήθως επιλέγεται ως τυπική κανονική κατανομή $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Η οριακή πιθανοφάνεια των δεδομένων μπορεί να εκφραστεί ως:

$$p_\theta(x) = \int p_\theta(x|z)p(z)dz. \tag{1.15}$$

Ωστόσο, αυτό το ολοκλήρωμα είναι γενικά αδύνατο να εκτιμηθεί. Οι ΠΑΚ παρακάμπτουν αυτή την πρόκληση εισάγοντας μια προσεγγιστική εκ των υστέρων κατανομή $q_\phi(z|x)$, παραμετροποιημένη από ένα νευρωνικό δίκτυο (κωδικοποιητή), και βελτιστοποιώντας ένα παραλλασσόμενο κάτω φράγμα στην λογαριθμική πιθανοφάνεια. Το **Κάτω Φράγμα Αποδείξεων** (ΚΦΑ, αγγλ. «Evidence Lower BOund», «ELBO») προκύπτει χρησιμοποιώντας την ανισότητα του Jensen:

$$\log p_\theta(x) \geq \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)\|p(z)) \tag{1.16}$$

$$= \mathcal{L}(\theta, \phi; x). \tag{1.17}$$

Το ΚΦΑ αποτελείται από δύο όρους: έναν όρο **ανακατασκευής** που ενθαρρύνει τον αποκωδικοποιητή να ανακατασκευάσει με ακρίβεια την είσοδο από την λανθάνουσα αναπαράσταση και έναν όρο **κανονικοποίησης** που περιορίζει την προσεγγιστική εκ των υστέρων

κατανομή ώστε να παραμείνει κοντά στην εκ των προτέρων κατανομή. Αυτή η διατύπωση επιτρέπει την ταυτόχρονη βελτιστοποίηση τόσο του γεννητικού μοντέλου $p_\theta(x|z)$ (αποκωδικοποιητής) όσο και του μοντέλου συμπερασμού $q_\phi(z|x)$ (κωδικοποιητής).

Το τέχνασμα της επαναπαραμετροποίησης. Μια κρίσιμη καινοτομία στους ΠΑΚ είναι το τέχνασμα της επαναπαραμετροποίησης [56], το οποίο επιτρέπει την οπισθοδιάδοση μέσω στοχαστικών κόμβων. Αντί να λαμβάνει δείγματα απευθείας από την $q_\phi(z|x)$, η μέθοδος εκφράζει την τυχαία μεταβλητή ως μια ντετερμινιστική συνάρτηση των παραμέτρων και μιας βοηθητικής μεταβλητής θορύβου. Για μια κανονική εκ των υστέρων κατανομή $q_\phi(z|x) = \mathcal{N}(\mu_\phi(x), \sigma^2_\phi(x))$, η επαναπαραμετροποίηση είναι:

$$z = \mu_\phi(x) + \sigma_\phi(x) \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{1.18}$$

όπου το $\odot$ δηλώνει το γινόμενο Hadamard (στοιχείο προς στοιχείο). Αυτή η μετατροπή επιτρέπει στις παραγώγους (gradients) να ρέουν μέσω της λειτουργίας δειγματοληψίας, επιτρέποντας τη βελτιστοποίηση ολόκληρου του πλαισίου χρησιμοποιώντας την κλασική οπισθοδιάδοση.

Μεθοδολογίες. Οι κυριότερες παραλλαγές των ΠΑΚ περιλαμβάνουν την εισαγωγή ενός παράγοντα β, καθώς και διάφορα σχήματα προγραμματισμού για την τιμή του, με στόχο τον έλεγχο της απόκλισης Kullback-Leibler και τη συνακόλουθη παρεμπόδιση της κατάρρευσης της εκ των υστέρων κατανομής [58, 59], τη χρήση πιο εξεζητημένων συναρτήσεων για την εκτίμηση της απώλειας ανακατασκευής [60, 61], καθώς και την εξεύρεση τρόπων για τη γεφύρωση του χάσματος εκφραστικότητας μεταξύ της προσεγγιστικής και της πραγματικής εκ των υστέρων κατανομής [62, 63].

Γεννητικά Ανταγωνιστικά Δίκτυα

Τα **Γεννητικά Ανταγωνιστικά Δίκτυα** (αγγλ. «Generative Adversarial Networks», «GANs») [64] έφεραν επανάσταση στη γεννητική μοντελοποίηση, εισάγοντας ένα πλαίσιο ανταγωνιστικής εκπαίδευσης που αντιπαραθέτει δύο νευρωνικά δίκτυα σε ένα παιχνίδι minimax. Σε αντίθεση με τους ΠΑΚ που βασίζονται σε ρητό υπολογισμό πιθανοφάνειας, τα ΓΑΔ μαθαίνουν να παράγουν ρεαλιστικά δείγματα μέσω ανταγωνισμού μεταξύ ενός **γεννητικού** δικτύου (generator) που παράγει συνθετικά δεδομένα και ενός **διαχωριστικού** δικτύου (discriminator) που διακρίνει μεταξύ πραγματικών και παραγόμενων δειγμάτων.

Θεωρητικά θεμέλια. Το πλαίσιο των ΓΑΔ διατυπώνει τη γεννητική μοντελοποίηση ως παιχνίδι minimax δύο παικτών. Το γεννητικό δίκτυο $G_\theta : \mathbb{R}^d \to \mathbb{R}^n$ απεικονίζει τυχαία διανύσματα θορύβου $z \sim p_z(z)$ στον χώρο δεδομένων, ενώ το διαχωριστικό δίκτυο $D_\phi : \mathbb{R}^n \to [0, 1]$ εκτιμά την πιθανότητα ένα δεδομένο δείγμα να προέρχεται από την πραγματική κατανομή δεδομένων και όχι από τη γεννήτρια. Ο στόχος της εκπαίδευσης είναι:

$$\min_\theta \max_\phi V(D_\phi, G_\theta) = \mathbb{E}_{x \sim p_{\text{data}}}[\log D_\phi(x)] + \mathbb{E}_{z \sim p_z}[\log(1 - D_\phi(G_\theta(z)))]. \tag{1.19}$$

Στη βέλτιστη λύση, η γεννήτρια μαθαίνει να προσεγγίζει την κατανομή δεδομένων p_{data} έτσι ώστε το διαχωριστικό δίκτυο να μην μπορεί να διακρίνει μεταξύ πραγματικών και παραγόμενων δειγμάτων, επιτυγχάνοντας $D^*(x) = 1/2$ για όλα τα x.

Μεθοδολογίες. Η εκπαίδευση των ΓΑΔ παρουσιάζει μοναδικές προκλήσεις λόγω της ανταγωνιστικής της φύσης. Το γεννητικό και το διαχωριστικό δίκτυο πρέπει να βρίσκονται σε ισορροπία - εάν το ένα δίκτυο γίνει πολύ ισχυρό σε σχέση με το άλλο, η εκπαίδευση μπορεί να γίνει ασταθής ή αναποτελεσματική. Για τον λόγο αυτό, έχουν αναπτυχθεί διάφορες μέθοδοι, όπως αντικατάσταση της απόκλισης Jensen-Shannon με την απόσταση Wasserstein [11] και επιβολή ποινής στις παραγώγους [65] για σταθερότερη εκπαίδευση, καθώς και ενσωμάτωση της τεχνικής μεταφοράς στιλ AdaIn [12] για μεγαλύτερο έλεγχο στα παραγόμενα δείγματα.

Μοντέλα διάχυσης

Τα μοντέλα διάχυσης [66, 67] είναι μια κατηγορία μοντέλων που βασίζονται στη θεωρία των στοχαστικών διαφορικών εξισώσεων (ΣτΔΕ) και μαθαίνουν να αντιστρέφουν μια διαδικασία σταδιακής εισαγωγής θορύβου [68]. Ο γεννητικός μηχανισμός εννοείται ως η χρονική αντιστροφή μιας διαδικασίας προς τα εμπρός διάχυσης που αλλοιώνει προοδευτικά τα δεδομένα έως ότου αυτά δεν μπορούν να διακριθούν πλέον από τυχαίο θόρυβο που ακολουθεί την κανονική κατανομή.

Θεωρητικά θεμέλια. Η διαδικασία της **προς τα εμπρός διάχυσης** ορίζεται ως μια μαρκοβιανή αλυσίδα που προσθέτει σταδιακά γκαουσιανό θόρυβο στο δείγμα x_0 σε T διακριτά χρονικά βήματα σύμφωνα με ένα ορισμένο πρόγραμμα διακύμανσης $\{\beta_t\}_{t=1}^T$:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}). \tag{1.20}$$

Μια βασική ιδιότητα αυτής της διαδικασίας είναι ότι η περιθώρια κατανομή σε οποιοδήποτε χρονικό βήμα t μπορεί να δειγματοληφθεί σε κλειστή μορφή χωρίς να διατρέχεται κάθε φορά η αλυσίδα. Αν θέσουμε $\alpha_t = 1 - \beta_t$ και $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$, προκύπτει:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}). \tag{1.21}$$

Αυτή η ιδιότητα είναι κρίσιμη για την αποδοτική εκπαίδευση, καθώς επιτρέπει την άμεση δειγματοληψία του x_t για οποιοδήποτε t μέσω της επαναπαραμετροποίησης: $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, όπου $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Η γεννητική διαδικασία, ή **προς τα πίσω διάχυση**, στοχεύει στην ανάκτηση των αρχικών δεδομένων μέσω της εκμάθησης των υπό συνθήκη κατανομών $p_\theta(x_{t-1}|x_t)$. Αυτές παραμετροποιούνται ως κανονικές κατανομές, των οποίων η μέση τιμή προβλέπεται από ένα νευρωνικό δίκτυο $\mu_\theta(x_t, t)$:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2\mathbf{I}). \tag{1.22}$$

Ξεκινώντας από καθαρό θόρυβο $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, το μοντέλο αποθορυβοποιεί επαναληπτικά το δείγμα για να παραγάγει μια τελική εκτίμηση.

Στόχος εκπαίδευσης. Ενώ το μοντέλο μπορεί να εκπαιδευτεί με τη μεγιστοποίηση του κατώτατου φράγματος διακύμανσης (αγγλ. «variational lower bound») στη λογαριθμική πιθανοφάνεια, έχει αποδειχθεί [66] ότι ένας απλοποιημένος στόχος, ο οποίος μπορεί να ερμηνευθεί ως απώλεια αποθορυβοποίησης μέσω αντιστοίχισης βαθμολογίας, αποδίδει ανώτερα αποτελέσματα. Ο στόχος είναι να εκπαιδευτεί ένα δίκτυο ϵ_θ με σκοπό την πρόβλεψη της συνιστώσας θορύβου ϵ από το θορυβώδες δείγμα x_t:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t\sim\mathcal{U}\{1,T\},x_0,\epsilon\sim\mathcal{N}(0,\mathbf{I})}\left[\|\epsilon-\epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0+\sqrt{1-\bar{\alpha}_t}\epsilon,t)\|^2\right]. \tag{1.23}$$

Διατύπωση βάσει βαθμολογίας. Το όριο συνεχούς χρόνου των μοντέλων διάχυσης παρέχει μια ενοποιημένη οπτική μέσω των ΣτΔΕ [69]. Η προς τα εμπρός διαδικασία μπορεί να εκφραστεί ως η λύση μιας ΣτΔΕ:

$$dx = f(x,t)dt + g(t)dw, \tag{1.24}$$

όπου $f(\cdot,t)$ είναι ένας συντελεστής ολίσθησης και $g(t)$ ένας συντελεστής διάχυσης. Η αντίστροφη αυτής της διαδικασίας είναι επίσης μια ΣτΔΕ που εξαρτάται από τη συνάρτηση βαθμολογίας της κατανομής των δεδομένων, $\nabla_x \log p_t(x)$:

$$dx = [f(x,t) - g(t)^2\nabla_x \log p_t(x)]dt + g(t)d\bar{w}. \tag{1.25}$$

Το νευρωνικό δίκτυο ϵ_θ εκπαιδεύεται για να προσεγγίζει αυτή τη συνάρτηση βαθμολογίας (όπως παρουσιάζεται στο [70]), καθώς σχετίζονται μέσω της σχέσης $\nabla_x \log p_t(x_t) \approx -\frac{\epsilon_\theta(x_t,t)}{\sqrt{1-\bar{\alpha}_t}}$.

Αυτό η οπτική που βασίζεται στη βαθμολογία φέρνει στο φως μια αντίστοιχη ντετερμινιστική διαδικασία γνωστή ως ΣΔΕ ροής πιθανότητας [69], η οποία μεταφέρει δείγματα από την κατανομή θορύβου στην κατανομή δεδομένων ακολουθώντας το διανυσματικό πεδίο που έχει μάθει. Η διατύπωση με ΣΔΕ είναι ιδιαίτερα χρήσιμη για δύο λόγους:

- **Γρήγορη δειγματοληψία**: Οι αριθμητικοί επιλυτές ΣΔΕ μπορούν να προσεγγίσουν την τροχιά με σημαντικά λιγότερα βήματα από τους στοχαστικούς δειγματολήπτες, παράγοντας δείγματα υψηλής ποιότητας σε μόλις 10-50 βήματα [71, 72].

- **Ελεγχόμενη παραγωγή**: Η ντετερμινιστική διαδρομή παρέχει μια μοναδική, αντιστρέψιμη λανθάνουσα αναπαράσταση για κάθε σημείο δεδομένων, η οποία είναι πολύτιμη για εργασίες όπως η επεξεργασία εικόνων και η παρεμβολή [73].

Η διατύπωση αυτή αναδιαμορφώνει τη διαδικασία δημιουργίας ως εκμάθηση ενός διανυσματικού πεδίου που καθοδηγεί τα δείγματα από μια απλή κατανομή θορύβου στην πολλαπλότητα (manifold) των δεδομένων. Η οπτική της εκμάθησης ενός διανυσματικού πεδίου για τη μεταφορά πυκνοτήτων πιθανότητας είναι ιδιαίτερα χρήσιμη και πρόσφατα γενικεύθηκε από την Αντιστοίχιση Ροής [74], μια τεχνική με εξέχοντα ρόλο στη μέθοδο που προτείνεται στο κεφάλαιο 4 αυτής της εργασίας.

Αντιστοίχιση Ροής

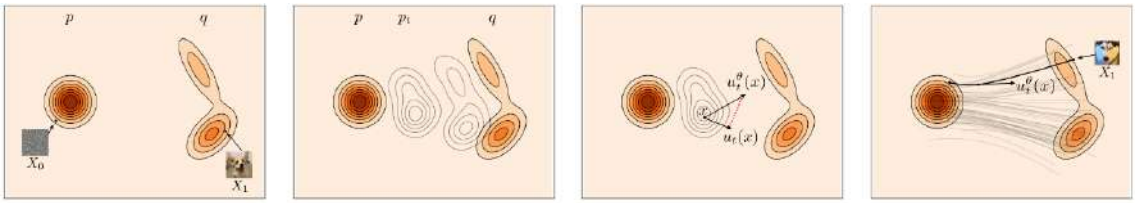
Η **Αντιστοίχιση Ροής (ΑΡ)** (αγγλ. «Flow Matching», «FM») [74, 5] αποτελεί σημαντική εξέλιξη στο πεδίο της γεννητικής μοντελοποίησης, διατυπώνοντας το πρόβλημα ως εκμάθηση διανυσματικών πεδίων που μετακινούν κατανομές πιθανοτήτων. Σε αντίθεση με τα μοντέλα διάχυσης που βασίζονται σε στοχαστικές διαδικασίες, η ΑΡ στοχεύει στην εκμάθηση ντετερμινιστικών ΣΔΕ, που πραγματοποιούν αντιστοίχιση μεταξύ των κατανομών.

Θεωρητικά θεμέλια. Η ΑΡ υποθέτει την ύπαρξη ενός χρονοεξαρτώμενου διανυσματικού πεδίου $v : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$, που υλοποιεί την αντιστοίχιση $v : (t,x) \mapsto v_t(x)$. Το πεδίο αυτό επάγει μια ροή $\phi : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$, που υλοποιεί την αντιστοίχιση $\phi : (t,x) \mapsto \phi_t(x)$. Η ροή ικανοποιεί την εξής σχέση:

$$\frac{d\phi_t(x)}{dt} = v_t(\phi_t(x)), \quad \phi_0(x) = x. \tag{1.26}$$

Η ιδιότητα της ροής είναι η μεταφορά δειγμάτων από μια απλή κατανομή πηγής p_0 (συνήθως κανονική) στη χρονική στιγμή $t = 0$ στην κατανομή στόχου p_1 στη χρονική στιγμή $t = 1$. Η εξέλιξη των πυκνοτήτων πιθανότητας ακολουθεί την εξίσωση συνέχειας:

$$\frac{\partial p_t}{\partial t} + \nabla \cdot (p_t v_t) = 0. \tag{1.27}$$



(a) Data. (b) Path design. (c) Training. (d) Sampling.

Figure 1.5. *Σκιαγράφηση της Αντιστοίχισης Ροής. (α) Ο στόχος είναι η εύρεση μιας ροής που αντιστοιχίζει δείγματα X_0 από μια γνωστή κατανομή πηγής ή μια κατανομή θορύβου p σε δείγματα X_1 από μια άγνωστη κατανομή δεδομένων q. (β) Για να το επιτύχουμε αυτό, σχεδιάζουμε ένα μονοπάτι πιθανότητας συνεχούς χρόνου $(p_t)_{0\le t\le 1}$ που παρεμβάλλεται μεταξύ των $p := p_0$ και $q := p_1$. (γ) Κατά την εκπαίδευση, χρησιμοποιούμε παλινδρόμηση για να εκτιμήσουμε το διανυσματικό πεδίο u_t που παράγει την p_t. (δ) Για να αντλήσουμε ένα νέο δείγμα από την κατανομή στόχου $X_1 \sim q$, ολοκληρώνουμε το εκτιμώμενο διανυσματικό πεδίο $u_t^\theta(X_t)$ από $t = 0$ έως $t = 1$, όπου το $X_0 \sim p$ είναι ένα νέο δείγμα της αρχικής κατανομής. Πηγή: [5]*

Αντιστοίχιση ροής υπό συνθήκη. Εφόσον η απευθείας εκμάθηση του περιθώριου διανυσματικού πεδίου είναι ανέφικτη, η ΑΡ κατασκευάζει ροές υπό συνθήκη $\phi_t(x|x_0, x_1)$ που παρεμβάλλονται μεταξύ του σημείου προέλευσης $x_0 \sim p_0$ και του σημείου προορισμού $x_1 \sim p_1$. Κατά συνέπεια, η διαδρομή και το αντίστοιχο διανυσματικό πεδίο γίνονται απλά και εύκολα να οριστούν. Για την ευρέως χρησιμοποιούμενη διαδρομή **Βέλτιστης Μεταφοράς (ΒΜ)**

(αγγλ. «Optimal Transport», «OT»), η ροή προκύπτει απλώς ως μια γραμμική παρεμβολή:

$$\phi_t(x_0, x_1) = (1-t)x_0 + tx_1, \tag{1.28}$$

με το αντίστοιχο υπό συνθήκη διανυσματικό πεδίο:

$$v_t(x_0, x_1) = \frac{d}{dt}\phi_t(x_0, x_1) = x_1 - x_0. \tag{1.29}$$

Το περιθώριο διανυσματικό πεδίο είναι τότε:

$$v_t(x) = \mathbb{E}_{(x_0,x_1)\sim\pi}[v_t(x|x_0, x_1)|\phi_t(x_0, x_1) = x], \tag{1.30}$$

όπου π είναι μια σύζευξη μεταξύ p_0 και p_1, δηλαδή μια από κοινού κατανομή πιθανότητας, της οποίας οι περιθώριες κατανομές στους αντίστοιχους άξονες είναι p_0 και p_1.

Στόχος εκπαίδευσης. Ο στόχος της εκπαίδευσης ελαχιστοποιεί την απόσταση ℓ_2 μεταξύ του διανυσματικού πεδίου $v_t^\theta(x)$ που έχει μάθει το μοντέλο και του διανυσματικού πεδίου στόχου $v_t(x_t|x_0, x_1)$ μέσω απλής παλινδρόμησης:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}[0,1],(x_0,x_1)\sim\pi}\left[\|v_t^\theta(x_t) - v_t(x_t|x_0, x_1)\|^2\right], \tag{1.31}$$

όπου $x_t = \phi_t(x_0, x_1)$. Στην περίπτωση της διαδρομής BM, η αντικατάσταση της εξ. 1.29 στην εξ. 1.31 δίδει:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}[0,1],(x_0,x_1)\sim\pi}\left[\|v_t^\theta(x_t) - (x_1 - x_0)\|^2\right]. \tag{1.32}$$

Στοχαστικοί παρεμβολείς. Για να εξασφαλιστεί αριθμητική ευστάθεια και να αντιμετωπιστούν πιθανές ιδιομορφίες, συχνά χρησιμοποιούνται **Στοχαστικοί Παρεμβολείς** (ΣΠ, αγγλ. «Stochastic Interpolants», «SIs») [75, 76]:

$$\phi_t(x_0, x_1) = (1-t)x_0 + tx_1 + \sigma_t\epsilon, \tag{1.33}$$

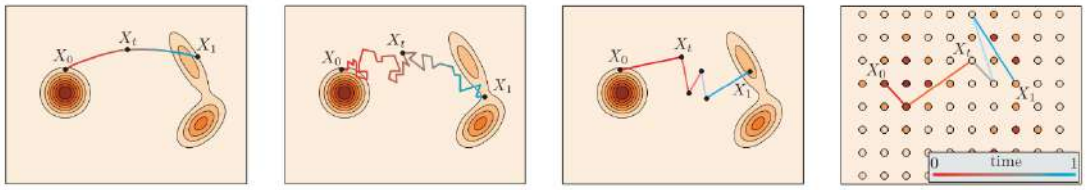
όπου $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ και σ_t είναι ένα πρόγραμμα θορύβου, συνήθως $\sigma_t = \sigma\min(t, 1-t)$ για κάποιο μικρό $\sigma > 0$. Το αντίστοιχο υπό συνθήκη διανυσματικό πεδίο γίνεται:

$$v_t(x_t|x_0, x_1) = \frac{x_1 - (1-\sigma_t)x_0 - \sigma_t x_t}{\sigma_t} + \frac{d\sigma_t}{dt}\epsilon. \tag{1.34}$$

Πλεονεκτήματα σε σχέση με συμβατικές μεθόδους:

1. **Ευστάθεια εκπαίδευσης**: Σε αντίθεση με τα ΓΑΔ, τα οποία απαιτούν προσεκτική εξισορρόπηση των ανταγωνιστικών απωλειών, η ΑΡ βελτιστοποιεί έναν απλό στόχο παλινδρόμησης.

2. **Ακριβής πιθανοφάνεια**: Η ΑΡ παρέχει ακριβή υπολογισμό πιθανοφάνειας χωρίς



(a) Flow (b) Diffusion (c) Jump (d) CTMC

Figure 1.6. *Τέσσερις χρονικά συνεχείς διαδικασίες $(X_t)_{0\leq t\leq 1}$ που μετακινούν το δείγμα προέλευσης X_0 σε ένα δείγμα προορισμού X_1. Αυτές είναι* **(α)** *μια ροή σε συνεχή χώρο κατάστασης,* **(β)** *μια διάχυση σε συνεχή χώρο κατάστασης,* **(γ)** *μια διαδικασία άλματος σε συνεχή χώρο κατάστασης (οι πυκνότητες απεικονίζονται ως ισοσταθμικές καμπύλες) και* **(δ)** *μια διαδικασία άλματος σε διακριτό χώρο καταστάσεων (οι καταστάσεις απεικονίζονται ως δίσκοι, οι πιθανότητες με χρώματα). Πηγή: [5]*

να απαιτεί υπολογιστικά επαχθείς εκτιμήσεις ιχνών πινάκων, σε αντίθεση με τις παραδοσιακές ροές κανονικοποίησης που βασίζονται στον τύπο αλλαγής μεταβλητών.

3. **Υψηλή ποιότητα δειγμάτων**: Η διατύπωση συνεχούς χρόνου επιτρέπει την ομαλή παρεμβολή μεταξύ κατανομών, με αποτέλεσμα σε αρκετές περιπτώσεις υψηλότερη ποιότητα δειγμάτων σε σύγκριση με τις μεθόδους διακριτού χρόνου.

4. **Υπολογιστική αποδοτικότητα**: Ο στόχος εκπαίδευσης με βάση την παλινδρόμηση καθώς και η διατύπωση με βάση τις ΣΔΕ επιφέρουν σημαντική επιτάχυνση, καθιστώντας την AP κατάλληλη για εφαρμογές μεγάλης κλίμακας.

1.3 Μεθοδολογία

1.3.1 CrossFlowDG

Η πρότερη ανασκόπηση της βιβλιογραφίας μάς οδηγεί με φυσικό τρόπο στην ανάπτυξη του *CrossFlowDG*, ενός πλαισίου που αντιμετωπίζει το τροπικό κενό χρησιμοποιώντας τρία βασικά στοιχεία: (1) μια *Τράπεζα Κειμενικών Τομέων* (TKT), δηλαδή μια λίστα με κειμενικές περιγραφές στιλιστικής φύσης, (2) μια *Τετραπλή Αντιπαραθετική Απώλεια* (TAA) για ενδοτροπική και διατροπική ευθυγράμμιση, και (3) μια μονάδα *Διατροπικής Αντιστοίχισης Ροής* (ΔΑΡ) που μαθαίνει μια ντετερμινιστική αντιστοιχία από εικόνες με πιθανή επίδραση του τομέα σε αναπαραστάσεις κειμένου ανεξάρτητες από τον τομέα. Μια επισκόπηση του πλαισίου απεικονίζεται στο Σχήμα 1.7.

Τράπεζα Κειμενικών Τομέων

Κατασκευάζουμε την **Τράπεζα Κειμενικών Τομέων (TKT)** ως ένα σύνολο $\mathcal{D} = \{d_1, d_2, \ldots, d_K\}$ από K περιγραφές πεδίου, ακολουθώντας το πρότυπο "a `{domain}` of a', όπου το `domain` περιλαμβάνει στιλιστικές παραλλαγές όπως "photograph"', "painting", "drawing" και παρόμοιους όρους που αποτυπώνουν διαφορετικά οπτικά στυλ. Κατά τη διάρκεια της εκπαίδευσης, κάθε εικόνα x_i με ετικέτα y_i συνδυάζεται με μια κειμενική περιγραφή που δημιουργείται με τυχαία δειγματοληψία από την TKT $d \sim \mathcal{U}(\mathcal{D})$ και συνενώνεται

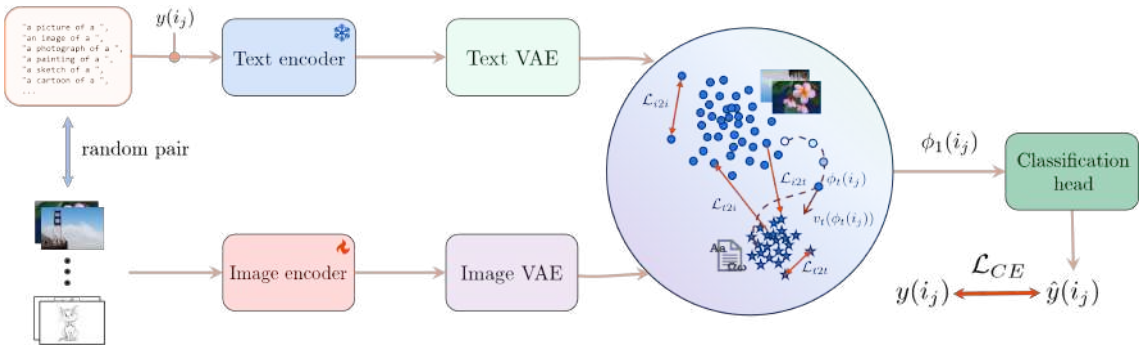


Figure 1.7. *Επισκόπηση του πλαισίου* FlowDG.

με το όνομα της κατηγορίας: $t_i = d + y_i$ ("+" όπως στη συνένωση συμβολοσειρών στην Python). Έτσι, παράγονται περιγραφές όπως "a photograph of a dog" ή "a sketch of a car". Είναι σημαντικό να σημειωθεί ότι ο η κειμενική περιγραφή επιλέγεται ανεξάρτητα για κάθε επανάληψη εκπαίδευσης, πράγμα που σημαίνει ότι η ίδια εικόνα μπορεί να αντιστοιχιστεί με διαφορετικές περιγραφές τομέα σε διαφορετικές εποχές.

Αυτή η τυχαία στρατηγική αντιστοίχισης εξυπηρετεί έναν διπλό σκοπό. Πρώτον, εκθέτοντας κάθε εικόνα σε πολλαπλές κειμενικές περιγραφές κατά τη διάρκεια της εκπαίδευσης, το μοντέλο μαθαίνει ότι το ίδιο οπτικό περιεχόμενο μπορεί να περιγραφεί με διάφορους τρόπους, ενθαρρύνοντας την εξαγωγή σημασιολογικών χαρακτηριστικών που δεν μεταβάλλονται ανάλογα με τον τομέα (συγκεκριμένα πληροφορίες κατηγορίας) και παραμένουν σταθερά σε όλες αυτές τις παραλλαγές. Δεύτερον, η κατανομή των ενσωματώσεων κειμένου που δημιουργούνται από την TKT για κάθε κατηγορία δημιουργεί έναν σημασιολογικά πλούσιο χώρο: αντί οι εικόνες να ευθυγραμμίζονται με μια μοναδική διανυσματική αναπαράσταση κειμένου ανά κλάση, ενθαρρύνονται να ευθυγραμμιστούν με μια ποικιλία κειμενικών αναπαραστάσεων.

Τετραπλή Αντιπαραθετική Απώλεια

Για να επιτύχουμε την εκμάθηση αναπαράστασεων ανεξάρτητων από τον εκάστοτε τομέα μέσω της ευθυγράμμισης με την TKT, χρησιμοποιούμε μια **Τετραπλή Αντιπαραθετική Απώλεια (TAA)** που αποτελείται από 4 συμπληρωματικούς όρους απώλειας: (1) εικόνα-σε-εικόνα ($\mathcal{L}_{i2i}$), για να δημιουργήσουμε συμπαγείς συστάδες εικόνων της ίδιας κλάσης, (2) κείμενο-σε-κείμενο ($\mathcal{L}_{t2t}$), με τον ίδιο στόχο όπως το (1) αλλά για τις αναπαραστάσεις κειμένου, και (3-4) διατροπικός ($\mathcal{L}_{i2t}$ και $\mathcal{L}_{t2i}$), για να γεφυρώσουμε το τροπικό κενό επιβάλλοντας την ευθυγράμμιση μεταξύ των αναπαραστάσεων εικόνων και κειμένου της ίδιας κλάσης.

Για να εφαρμόσουμε τον αντιπαραθετικό στόχο, προβάλλουμε τα χαρακτηριστικά που εξαγάγαμε σε έναν Ευκλείδειο κοινό λανθάνοντα χώρο χρησιμοποιώντας δύο ξεχωριστούς ΠΑΚ, έναν για κάθε τροπικότητα. Ας υποθέσουμε ότι f^i είναι ο προεκπαιδευμένος κωδικοποιητής εικόνας, g^i ο ΠΑΚ εικόνας, f^t ο προεκπαιδευμένος κωδικοποιητής κειμένου και g^t ο ΠΑΚ κειμένου. Το $z_j^i = g^i(f^i(x_j))$ δηλώνει τη λανθάνουσα αναπαράσταση της εικόνας και το $z_j^t = g^t(f^t(t_j))$ την αντίστοιχη αναπαράσταση του κειμένου. Για τον υπολογισμό της απώλειας, τα χαρακτηριστικά κανονικοποιούνται στη μοναδιαία υπερσφαίρα μέσω της h, έτσι

ώστε $\|z_j\|_2 = 1$. Ορίζουμε την ομοιότητα συνημιτόνου ως:

$$s(z_a, z_b) = \frac{h(z_a)^\top h(z_b)}{\tau}, \tag{1.35}$$

όπου $\tau > 0$ είναι μια παράμετρος θερμοκρασίας. Κάθε όρος αντιπαραθετικής απώλειας μεταξύ των τροπικοτήτων k και l ορίζεται ως:

$$\mathcal{L}_{k2l} = -\frac{1}{N}\sum_{j=1}^{N} \log \frac{\exp(s(z_j^k, z_j^l))}{\sum_{m=1}^{N} \exp(s(z_j^k, z_m^l))}. \tag{1.36}$$

Η συνολική απώλεια είναι:

$$\mathcal{L}_{\text{FC}} = \mathcal{L}_{i2t} + \mathcal{L}_{t2i} + \frac{1}{2}(\mathcal{L}_{t2t} + \mathcal{L}_{i2i}). \tag{1.37}$$

Αυτός ο στόχος ενθαρρύνει την διατροπική ευθυγράμμιση ανά κατηγορία, διατηρώντας παράλληλα τη συνοχή εντός συστάδας.

Διατροπική Αντιστοίχιση Ροής

Όπως αναφέρθηκε προηγουμένως, η τυπική αντιπαραθετική μάθηση ομοιότητας συνημιτόνου δεν καλύπτει πλήρως το τροπικό κενό. Επομένως, χρησιμοποιούμε τη **Διατροπική Αντιστοίχιση Ροής (ΔΑΡ)** χωρίς θόρυβο, με στόχο την εκμάθηση ενός συνεχούς μετασχηματισμού που αντιστοιχίζει τις λανθάνουσες εικόνες στις λανθάνουσες κειμενικές αναπαραστάσεις, διατηρώντας παράλληλα τη γεωμετρία του χώρου. Σε αντίθεση με τις συμβατικές μεθόδους βασισμένες σε ροές που αντιστοιχίζουν θόρυβο σε δεδομένα, υιοθετούμε μια ντετερμινιστική διατροπική ροή που συνδέει τις πολλαπλότητες εικόνων και κειμένων στον λανθάνοντα χώρο, παρόμοια με το [77].

Δεδομένων των συζευγμένων λανθανουσών αναπαραστάσεων εικόνων-κειμένων $(z_{\text{img}}, z_{\text{txt}})$, ορίζουμε ένα παρεμβαλλόμενο δείγμα στο χρόνο $t \in [0, 1]$:

$$z_t = (1-t)z_{\text{img}} + t z_{\text{txt}}. \tag{1.38}$$

Το πραγματικό πεδίο ταχυτήτων είναι:

$$v^*(z_t, t) = z_{\text{txt}} - z_{\text{img}}. \tag{1.39}$$

Ένα νευρωνικό μοντέλο ροής που προβλέπει την ταχύτητα $v^\theta(z_t, t)$ εκπαιδεύεται μέσω της:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}[0,1],(z_{\text{img}},z_{\text{txt}})\sim\pi_{\text{rand}}}\left[\|v_t^\theta(z_t) - v_t^*(z_t|z_{\text{img}}, z_{\text{txt}})\|^2\right], \tag{1.40}$$

όπου π_{rand} είναι η τυχαία σύζευξη εικόνας-κειμένου σε επίπεδο δέσμης (mini-batch) που περιγράφεται στο 4.1.1.

Κατά τον συμπερασμό, η ροή ορίζει μια ντετερμινιστική αντιστοίχιση από την εικόνα στο κείμενο:

$$z_1 = z_{\text{img}} + \int_0^1 v^\theta(z_t, t)\, dt, \tag{1.41}$$

η οποία προσεγγίζεται χρησιμοποιώντας έναν μικρό αριθμό διακριτών βημάτων:

$$z_1 \approx z_{\text{img}} + \sum_{i=0}^{T-1} v^{\theta}(z_{\frac{i}{T}}, \frac{i}{T})\frac{1}{T}. \tag{1.42}$$

1.3.2 OT-CrossFlowDG

Κίνητρο. Παρά τα πλεονεκτήματα του *CrossFlowDG*, ο στόχος μας να βελτιώσουμε την αποδοτικότητα του συμπερασμού και, ως εκ τούτου, να διευκολύνουμε την εφαρμογή σε συσκευές με περιορισμένες υπολογιστικές δυνατότητες, μας οδηγεί προς την κατεύθυνση του μοντέλου ροής. Πρέπει να σημειωθεί ότι, στην παρούσα υλοποίηση, το *CrossFlowDG* χρησιμοποιεί 12 διακριτά χρονικά βήματα για την προσέγγιση της ολοκλήρωσης της ΣΔΕ, πράγμα που σημαίνει ότι το μοντέλο ροής πρέπει να ερωτηθεί 12 φορές πριν δώσει την τελική αναπαράσταση τη χρονική στιγμή $t = 1$. Σε μια προσπάθεια να μειώσουμε τον αριθμό των βημάτων ολοκλήρωσης, σχεδιάζουμε τον Αλγόριθμο 4.2 για να διορθώσουμε την αφελή, τυχαία σύζευξη εικόνας-κειμένου και να εκκινήσουμε τη διατύπωση της AP υπό συνθήκη με ζεύγη που βασίζονται στη βέλτιστη μεταφορά, παρόμοια με το [78].

Αλγόριθμος σύζευξης βάσει Sinkhorn-Knopp ανά κατηγορία. Ο Αλγόριθμος 4.2 έχει σχεδιαστεί για να υπολογίζει την απόσταση βέλτιστης μεταφοράς μεταξύ ενός συνόλου αναπαραστάσεων εικόνων $\mathbf{E}_I \in \mathbb{R}^{B\times D}$ και κειμένου $\mathbf{E}_T \in \mathbb{R}^{K\times D}$, επιβάλλοντας παράλληλα έναν αυστηρό περιορισμό ευθυγράμμισης βάσει κατηγορίας. Αρχικά, υπολογίζεται ο πίνακας κόστους $\mathbf{C}$ χρησιμοποιώντας την **τετραγωνισμένη Ευκλείδεια απόσταση** μεταξύ όλων των ζευγών εικόνας-κειμένου μιας δέσμης (mini-batch), που σημαίνει ότι $B = K$ στην περίπτωσή μας. Έπειτα, ένα μεγάλο κόστος ποινής L εφαρμόζεται σε όλα τα ζεύγη διαφορετικών κλάσεων, επιτρέποντας τη σύζευξη μόνο μεταξύ αναπαραστάσεων εικόνας-κειμένου που ανήκουν στην ίδια κατηγορία. Ο προκύπτων πίνακας κόστους $\mathbf{C}_{\text{mask}}$ δίνεται ως είσοδος στον αλγόριθμο **Sinkhorn-Knopp** [79] (που περιγράφεται στον Αλγόριθμο 4.1 για ευκολία αναφοράς), ο οποίος αποδίδει το κατά προσέγγιση βέλτιστο πλάνο μεταφοράς $\mathbf{P}$. Η τελική βαθμωτή απώλεια $\mathcal{L}_{\text{CW}}$ ορίζεται ως το προσδοκώμενο κόστος σύμφωνα με το πλάνο $\mathbf{P}$ αθροισμένο μόνο πάνω στα έγκυρα ζεύγη. Μια βασική επιπλέον έξοδος αυτού του αλγορίθμου είναι το σύνολο των **βαρυκεντρικών στόχων** $\mathbf{B} \in \mathbb{R}^{K\times D}$, οι οποίοι υπολογίζονται ως ο σταθμισμένος μέσος όρος των αναπαραστάσεων κειμένου σύμφωνα με το πλάνο μεταφοράς $\mathbf{P}$.

1.4 Πειράματα

1.4.1 Σύνολα αξιολόγησης

Αξιολογούμε την ακρίβεια ταξινόμησης σε τέσσερα συνήθη σύνολα δεδομένων ΓΤ, συγκεκριμένα στα **TerraIncognita** [80], **PACS** [81], **VLCS** [82] και **OfficeHome** [83], υιοθετώντας το πρωτόκολλο αξιολόγησης *leave-one-domain-out*: εάν το σύνολο δεδομένων περιέχει N τομείς $\mathcal{D}_i, i \in \{1, 2, \ldots, N\}$, εκπαιδεύουμε σε $N-1$ τομείς, αξιολογούμε στον εναπομείναντα, επαναλαμβάνουμε για κάθε τομέα στόχου και υπολογίζουμε τον μέσο όρο

ως:

$$\text{LODO}_{\mathcal{D}} = \frac{1}{N}\sum_{i=1}^{N} R_{\setminus i}, \tag{1.43}$$

όπου $R_{\setminus i}$ δηλώνει την ακρίβεια του μοντέλου όταν εκπαιδεύεται στο $\mathcal{D} \setminus \{\mathcal{D}_i\}$.

Για δίκαιη σύγκριση με προηγούμενες μεθόδους, πραγματοποιούμε εκπαίδευση για 50 εποχές με 200 ενημερώσεις παραμέτρων ανά εποχή (10.000 επαναλήψεις συνολικά). Η αναφερόμενη ακρίβεια για το *CrossFlowDG* υπολογίζεται ως μέσος όρος 3 διαφορετικών τυχαίων σπόρων (seeds) μαζί με την τυπική απόκλιση, σε αντίθεση με κάποιες προηγούμενες σχετικές δουλειές. Λόγω περιορισμών πόρων, για το *OT-CrossFlowDG* εκτελούμε 3 δοκιμές, καθεμία με διαφορετικό αριθμό βημάτων ολοκλήρωσης της ΣΔΕ, για να διαπιστώσουμε κάποια πιθανή διαφορά στην επίδοση.

Table 1.1. *Μια ποσοτική θεώρηση των συνόλων δεδομένων που χρησιμοποιούνται για την αξιολόγηση.*

Σύνολο δεδομένων	**Τομέας**	**# εικόνων**	**# εικόνων**	**# κλάσεων**
TerraIncognita	L100	4741	24330	10
	L38	9736		
	L43	3970		
	L46	5883		
PACS	Art painting	2048	9991	7
	Cartoon	2344		
	Photo	1670		
	Sketch	3929		
VLCS	Caltech101	1415	10729	5
	LabelMe	2656		
	SUN09	3282		
	VOC2007	3376		
OfficeHome	Art	2427	15588	65
	Clipart	4365		
	Product	4439		
	Real	4357		

Για να πραγματοποιήσουμε μια εμπεριστατωμένη αξιολόγηση, συγκρίνουμε το προτεινόμενο μοντέλο μας με διάφορες αρχιτεκτονικές κορυφαίων επιδόσεων στο πλαίσιο της ΓΤ. Συγκεκριμένα, επιλέγουμε μεθόδους βασισμένες σε ResNet-50 [84], DeiT-S [85] και VMamba-T [9] με παρόμοιο αριθμό παραμέτρων και χρησιμοποιούμε την ευρέως διαδεδομένη προεκπαίδευση στο ImageNet-1K [86] για δίκαιη σύγκριση.

1.4.2 Λεπτομέρειες υλοποίησης

Στόχος εκπαίδευσης

Ο συνολικός στόχος εκπαίδευσης του *CrossFlowDG* συνδυάζει τους επιμέρους όρους απώλειας των ΠΑΚ, ΤΑΑ και ΔΑΡ, όπως ορίζονται στην υποενότητα 1.3.1:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{VAE}}^{\text{img}} \mathcal{L}_{\text{VAE}}^{\text{img}} + \lambda_{\text{VAE}}^{\text{txt}} \mathcal{L}_{\text{VAE}}^{\text{txt}} + \lambda_{\text{FC}} \mathcal{L}_{\text{FC}} + \lambda_{\text{FM}} \mathcal{L}_{\text{FM}}, \tag{1.44}$$

όπου $\mathcal{L}_{\text{VAE}}^{k} = \mathcal{L}_{\text{recon}}^{k} + \mathcal{L}_{\text{KL}}^{k}$ η απώλεια του ΠΑΚ που αντιστοιχεί στην τροπικότητα k και περιλαμβάνει έναν όρο ανακατασκευής (π.χ. απώλεια μέσου τετραγωνισμένου σφάλματος) και έναν όρο ομαλοποίησης (π.χ. απόκλιση Kullback-Leibler μεταξύ της κατανομής της λανθάνουσας αναπαράστασης και μιας τυπικής κανονικής κατανομής), ενώ οι συντελεστές λ_i ελέγχουν τις συνεισφορές των αντίστοιχων όρων. Για το *OT-CrossFlowDG*, το $\mathcal{L}_{\text{FC}}$ αντικαθίσταται απλά από το $\mathcal{L}_{\text{CW}}$, όπως περιγράφεται στην υποενότητα 1.3.2.

Όλα τα εκπαιδεύσιμα στοιχεία των μεθόδων μας εκπαιδεύονται από κοινού. Επίσης, όλα τα εκπαιδεύσιμα στοιχεία εκπαιδεύονται από το μηδέν, με εξαίρεση τον προεκπαιδευμένο κωδικοποιητή εικόνων.

Αρχιτεκτονική δικτύου

Σημείωση: Το «M» παρακάτω σημαίνει «εκατομμύρια».

- **Κωδικοποιητής εικόνας:** VMamba-T [9], 29M εκπαιδεύσιμες παράμετροι.
- **Κωδικοποιητής κειμένου:** Κωδικοποιητής κειμένου του CLIP [49], χωρίς εκπαιδεύσιμες παραμέτρους.
- **ΠΑΚ:** Κωδικοποιητής και αποκωδικοποιητής Multi-Layer Perceptron (MLP), χρησιμοποιήθηκαν 2 ΠΑΚ, με 4M και 1M εκπαιδεύσιμες παραμέτρους.
- **Μοντέλο ροής:** ResNet [84], 1M εκπαιδεύσιμες παράμετροι.
- **Κεφαλή ταξινόμησης:** MLP, 0,02M εκπαιδεύσιμες παράμετροι.

Η μέθοδος που προτείνουμε περιλαμβάνει συνολικά 36M εκπαιδεύσιμες παραμέτρους.

Λανθάνων χώρος

- **Λανθάνουσα διάσταση:** 256
- **Βήματα ολοκλήρωσης ΣΔΕ:** *CrossFlowDG*: 12, *OT-CrossFlowDG*: 1, 6, 12

Λεπτομέρειες εκπαίδευσης

- **Μέγεθος δέσμης:** 16 (ανά τομέα εκπαίδευσης)
- **Επαναλήψεις:** 10.000 συνολικές επαναλήψεις (50 εποχές × 200 ενημερώσεις παραμέτρων / εποχή)
- **Υλικό:** 1 × A10G GPU

1.4.3 Αποτελέσματα

Οι πίνακες των αποτελεσμάτων για τα δύο προτεινόμενα πλαίσια (*CrossFlowDG* και *OT-CrossFlowDG*) παρατίθενται στην ενότητα 5.4.

1.5 Συζήτηση

1.5.1 Συμπεράσματα

Σε αυτή την εργασία, παρουσιάζουμε δύο καινοτόμες μεθόδους για την αντιμετώπιση του κρίσιμου προβλήματος της ταξινόμησης εικόνων υπό συνθήκες μετατόπισης τομέα: το *CrossFlowDG* και το *OT-CrossFlowDG*, με το δεύτερο να αποτελεί μια παραλλαγή του πρώτου βασισμένη στη βέλτιστη μεταφορά (BM).

Το πρώτο προτεινόμενο πλαίσιο, ονόματι *CrossFlowDG*, πραγματοποιεί μια πρωτογενή ευθυγράμμιση εικόνας-κειμένου μέσω μιας αντιπαραθετικής απώλειας με βάση την ομοιότητα συνημιτόνου σε έναν Ευκλείδειο κοινό λανθάνοντα χώρο, ακολουθούμενη από τον μηχανισμό Αντιστοίχισης Ροής (AP), ο οποίος αντιμετωπίζει το τροπικό κενό μέσω της εκμάθησης ενός συνεχούς μετασχηματισμού μεταξύ των μερικώς ευθυγραμμισμένων αναπαραστάσεων εικόνας και κειμένου της ίδιας κατηγορίας.

Το *CrossFlowDG* επιτυγχάνει κορυφαία αποτελέσματα σε διάφορους τομείς των συνόλων δεδομένων αξιολόγησης, κυρίως στο TerraIncognita, το πιο απαιτητικό κατά την άποψή μας, τόσο ποσοτικά όσο και ποιοτικά. Η υψηλή δυσκολία του αποδεικνύεται από τις σταθερά χαμηλότερες τιμές ακρίβειας που επιτυγχάνονται σε σχέση με προηγούμενες μεθόδους, οι οποίες μόλις ξεπερνούν το 54% όπως φαίνεται στον Πίνακα 5.2. Ποιοτικά, οι εικόνες παρουσιάζουν σοβαρές μετατοπίσεις τομέα που χαρακτηρίζονται από θόλωμα λόγω κίνησης και συνθήκες χαμηλού φωτισμού, οι οποίες εμποδίζουν σε μεγάλο βαθμό την ακριβή ταξινόμηση ακόμη και από ανθρώπους.

Αντίστροφα, τα σχεδόν τέλεια αποτελέσματα σε ορισμένους τομείς υπογραμμίζουν τα όρια της βελτίωσης στην πράξη. Σε όλες τις προηγούμενες μεθόδους, ο τομέας «photo» του PACS και ο τομέας «Caltech101» του VLCS παρουσιάζουν εξαιρετικά κορεσμένη απόδοση, με ακρίβεια που υπερβαίνει το 99,0% στις πιο πρόσφατες μεθόδους. Αυτός ο υψηλός κορεσμός πιθανώς ανταναικλά το γεγονός ότι οι προεκπαιδευμένοι κωδικοποιητές έχουν συναντήσει (κατά την προεκπαίδευση) δείγματα παρόμοια (ή ακόμη και πανομοιότυπα) με αυτά που περιέχονται στο PACS και το VLCS. Κατά συνέπεια, οι μικρές διαφορές στην επίδοση σε αυτούς τους τομείς είναι πιθανώς λιγότερο ενδεικτικές της πραγματικής ικανότητας γενίκευσης του μοντέλου.

Επιπλέον, το *CrossFlowDG* παρουσιάζει υποβέλτιστη επίδοση στο OfficeHome και υποθέτουμε ότι ο κύριος λόγος για αυτή τη συμπεριφορά είναι ο μεγάλος αριθμός των κλάσεων του συνόλου αυτού. Συγκεκριμένα, το OfficeHome περιέχει 65 κατηγορίες, ένας παράγοντας που θεωρούμε πως εμποδίζει σημαντικά τη δημιουργία διακριτών συστάδων κατηγοριών στον λανθάνοντα χώρο. Αυτή η έλλειψη σαφώς διαχωρισμένων συστάδων μειώνει στη συνέχεια την αποτελεσματικότητα του μηχανισμού AP, ο οποίος βασίζεται ουσιαστικά στην εκμάθηση μιας συνεχούς διαδρομής μεταξύ της κατανομής εικόνων και της σημασιολογικά όμοιας κατανομής κειμένου.

Το δεύτερο προτεινόμενο πλαίσιο, ονόματι *OT-CrossFlowDG*, δίνει έμφαση στον αποδοτικότερο συμπερασμό, με τις βασικές διαφορές από το *CrossFlowDG* να είναι δύο: (1) η διατροπική απώλεια υπολογίζεται χρησιμοποιώντας την εντροπικά ομαλοποιημένη απόσταση 2-Wasserstein, όπως υπολογίζεται απο τον αλγόριθμο Sinkhorn-Knopp, (2) ο οποίος επιπλέον αξιοποιείται για την εξαγωγή βαρυκεντρικών στόχων από ένα κατά προσέγγιση βέλτιστο σχέδιο μεταφοράς. Αυτή η σύζευξη με βάση τη BM παράγει ένα βελτιωμένο σήμα επίβλεψης που μειώνει σημαντικά τον αριθμό των βημάτων ολοκλήρωσης που απαιτούνται από τον μηχανισμό της AP κατά τη διάρκεια του συμπερασμού.

Σχετικά με τα πειραματικά αποτελέσματα του *OT-CrossFlowDG*, παρατηρούμε ότι το OfficeHome αποτελεί μια αξιοσημείωτη εξαίρεση στην τάση της επίτευξης της υψηλότερης επίδοσης με λιγότερα βήματα προσέγγισης του ολοκλρώματος. Όταν ο λανθάνων χώρος καλείται να φιλοξενήσει τις 65 κατηγορίες του OfficeHome, οι προκύπτουσες συστάδες των κλάσεων γίνονται πυκνές και ενδεχομένως αλληλεπικαλύπτονται. Για την ακριβή πλοήγηση σε αυτόν τον πολύπλοκο χώρο, είναι πιθανό να απαιτείται μια λεπτομερέστερη διακριτοποίηση της τροχιάς, δηλαδή ένας μεγαλύτερος αριθμός βημάτων ολοκλήρωσης, προκειμένου να διατηρηθεί η ακρίβεια της ταξινόμησης.

1.5.2 Περιορισμοί και μελλοντική έρευνα

Παρά τα πλεονεκτήματα των μεθόδων που προτείνουμε, αναγνωρίζουμε ορισμένους βασικούς περιορισμούς που μπορούν να χρησιμεύσουν σαν πυξίδα για μελλοντική έρευνα:

Συνδυασμός σταθερότητας κατά την εκπαίδευση με αποδοτικότητα κατά τον συμπερασμό. Το *CrossFlowDG* χαίρει σταθερής εκπαίδευσης λόγω της χρήσης ομοιότητας συνημιτόνου, ενώ το *OT-CrossFlowDG* μπορεί να είναι υπολογιστικά ελαφρύτερο κατά τη διάρκεια του συμπερασμού, αλλά τείνει να υποφέρει από εξαφανιζόμενες ή ασταθείς κλίσεις, ίσως λόγω της υψηλής ευελιξίας που εισάγουν οι Ευκλείδειες αποστάσεις στον πίνακα κόστους. Μελλοντικές εργασίες θα μπορούσαν να συνδυάσουν τα πλεονεκτήματα και των δύο προσεγγίσεων, χρησιμοποιώντας αποστάσεις με βάση το συνημίτονο στο πλαίσιο της BM.

Γενικότητα των προτεινόμενων πλαισίων. Αξιολογήσαμε τις μεθόδους μας χρησιμοποιώντας μόνο το VMamba-T και τον κωδικοποιητή κειμένου CLIP ως προεκπαιδευμένα μοντέλα. Για το λόγο αυτό, απαιτούνται περαιτέρω πειράματα με διάφορους συνδυασμούς κωδικοποιητών εικόνας και κειμένου, προς ενίσχυση του επιχειρήματος (εμπειρικά) περί γενικότητας των πλαισίων μας. Αξίζει να σημειώσουμε πως ένα βασικό χαρακτηριστικό της προσέγγισής μας είναι ότι δεν βασίζεται σε τροπικότητες που έχουν ευθυγραμμιστεί ήδη από το στάδιο της προεκπαίδευσης, αλλά επιβάλλει ρητά την ευθυγράμμιση κατά τη διάρκεια της εκπαίδευσης. Κατά συνέπεια, το *CrossFlowDG* και το *OT-CrossFlowDG* θα μπορούσαν να επεκταθούν σε σενάρια που περιλαμβάνουν έναν αυθαίρετο αριθμό τροπικοτήτων, επιτρέποντας μια πιο ευέλικτη διατροπική ευθυγράμμιση.

Αξιολόγηση σε κορεσμένα σύνολα δεδομένων. Όπως αναφέρθηκε, ορισμένοι τομείς παρουσιάζουν κορεσμό της επίδοσης με τιμές άνω του 99%. Τέτοιου είδους αποτελέσ-

ματα ενδέχεται να αντανακλούν διαρροή δεδομένων από την προεκπαίδευση και όχι πραγματική ικανότητα ΓΤ. Μελλοντικές εργασίες θα πρέπει να δώσουν προτεραιότητα στην αξιολόγηση σε πιο απαιτητικά σύνολα που αντανακλούν καλύτερα τις μεταβολές του τομέα και αποφεύγουν τα φαινόμενα κορεσμού, π.χ. σύνολα δεδομένων με μεγαλύτερη ποικιλομορφία των τομέων ή που έχουν κατασκευαστεί έτσι ώστε να διαφέρουν από τα κοινώς χρησιμοποιούμενα σύνολα προεκπαίδευσης.

Chapter 2

Introduction

2.1 Problem statement

In an era where machine learning systems are increasingly deployed across diverse real-world environments, maintaining performance in spite of these variations is necessary for safe and reliable artificial intelligence. Domain Generalization (DG) involves assessing this ability of models to perform effectively in unseen target domains, but without access to target data during training [1, 13], and is therefore essential for practical applications ranging from autonomous vehicles navigating varied weather conditions and geographical locations [87, 88, 89], to medical diagnostic systems that must generalize across different patient populations and imaging protocols [90, 91, 92, 93].

Throughout the years, the research community has developed diverse approaches to tackle DG, which can be broadly categorized into several paradigms [1, 13]. Data augmentation strategies attempt to expand the training distribution through sophisticated transformations and synthesis techniques either in the input or feature space, while domain-invariant representation learning seeks to extract features that remain stable across domain shifts through adversarial training or explicit invariance constraints. Meta-learning approaches frame DG as learning to learn, enabling models to quickly adapt to new domains by simulating domain shifts during training. Ensemble methods use multiple models or domain-specific components to capture diverse domain characteristics, and regularization techniques explicitly penalize domain-specific features to promote generalization. More recently, multimodal approaches have gained significant attention by exploiting the complementary nature of different data modalities, such as vision, audio, and language, to learn more robust representations. The underlying hypothesis is that cross-modal information provides richer semantic grounding, potentially offering invariant signals that can mitigate unimodal domain shifts.

However, some multimodal approaches face a critical limitation for DG. While text embeddings serve as natural domain-invariant anchors, standard contrastive learning methods produce a phenomenon known as the *modality gap*, where the support of the image and text distributions are completely disjoint [3]. This disjointness is problematic when enforcing domain invariance involves regularizing image representations toward the domain-invariant text manifold. Consequently, without an explicit bridging mechanism, the domain-invariant knowledge encoded in text embeddings cannot be effectively ex-

ploited. In the context of DG, the modality gap poses a fundamental question:

Can bridging the modality gap
enhance domain generalization capabilities?

This motivates an exploration of flow matching (FM) [74, 5] as a potential solution. FM is a continuous normalizing flow framework that learns to transform between probability distributions through ordinary differential equations (ODEs). Unlike traditional generative models, FM provides explicit control over the transformation pathway, enabling direct interpolation and alignment between distributions. In the context of multimodal learning, FM can be leveraged to learn smooth transport maps between modality-specific representation spaces, potentially bridging the modality gap by constructing continuous trajectories that connect semantically equivalent embeddings across modalities. Furthermore, by learning these transport dynamics on diverse source domains, FM may capture domain-invariant transformation patterns that generalize to unseen target domains. This thesis investigates whether FM can serve as an effective bridge across the modality gap to enhance domain generalization, examining how its continuous transport framework can assist in learning domain-invariant features that enable generalization to new deployment environments.

2.2 Contributions

The main contributions of this thesis are summarized as follows:

1. We address the challenge of multimodal feature misalignment by introducing noise-free flow matching to bridge the modality gap between image and text embeddings.
2. We propose *CrossFlowDG*, a novel DG framework that combines cosine similarity contrastive learning for preliminary cross-modal alignment with flow matching to bridge the residual modality gap. Furthermore, our framework does not assume alignment during pretraining, unlike most prior work. Using the VMamba backbone, *CrossFlowDG* achieves state-of-the-art performance across multiple domains while maintaining inference efficiency.
3. We also introduce *OT-CrossFlowDG*, a *FlowDG* variant that incorporates a custom class-wise optimal transport algorithm to derive approximate optimal couplings between image embeddings and barycentric target text embeddings, which serve as refined, distributionally-aware supervision points to bootstrap the flow matching process. This superior guidance enables *OT-CrossFlowDG* to achieve its peak performance using only 1 inference step, making it more suitable for resource-constrained deployment.

2.3 Structure of this thesis

Chapter 3 contains a comprehensive literature review on the theory and applications relevant to our proposed frameworks, which are presented in Chapter 4. Details about

the benchmarks and the implementation as well as experimental results are included in Chapter 5, while in Chapter 6 we draw conclusions and discuss potential limitations and directions for future research.

Chapter 3

Theoretical background

Preliminary definitions

Artificial Intelligence (AI) is the broad field of computer science dedicated to creating systems capable of performing tasks that typically require human intelligence, such as decision-making, speech recognition, and visual perception [94]. AI is often categorized into systems that think or act rationally and those that think or act like humans.

Machine Learning (ML) is a subfield of AI focused on developing algorithms that allow computers to improve their performance on a specific task (T) with experience (E), as measured by some performance metric (P) [95]. ML systems learn patterns directly from data rather than being explicitly programmed with rules.

ML algorithms are broadly categorized based on the nature of the data and the feedback mechanism used during training:

- **Supervised learning (SL)**. It is the most common paradigm in machine learning. It involves training a model on a labeled dataset, where each training example is paired with a corresponding output label or value [55]. The goal of the model is to learn a mapping function from the input data to the output labels, allowing it to accurately predict the labels for new, unseen data. Tasks such as classification (predicting a categorical label) and regression (predicting a continuous value) fall under this category.

- **Unsupervised learning (UL)**. In contrast to SL, UL deals with unlabeled data. The algorithms are tasked with discovering hidden patterns, structures, or relationships within the data on their own [96]. This is often used for tasks such as clustering (grouping similar data points together) and dimensionality reduction (compressing data into a lower-dimensional representation while preserving its essential information).

- **Reinforcement learning (RL)**. A ML paradigm in which an *agent* (*i.e.*, "one who acts") learns to make a sequence of decisions by interacting with an *environment*. The agent receives a *reward* signal for taking correct actions and a penalty (or no reward) for incorrect ones. The objective is to learn an *optimal policy*—a set of actions that maximize the cumulative reward over time [97]. This approach is widely used in robotics, game-playing, and autonomous systems.

Deep Learning (DL) is a specialized subfield of Machine Learning that uses artificial neural networks with multiple hidden layers—hence the term "deep" [54]. The primary advantage of DL models is their ability to automatically learn hierarchical representations of data. For example, when processing an image, the initial layers of a deep neural network might learn to detect simple features like edges and curves, while subsequent layers combine these features to recognize more complex concepts like objects and faces. This hierarchical feature extraction makes DL particularly effective for complex tasks involving unstructured data like images, audio, and text.

Computer Vision (CV) is an interdisciplinary field that seeks to enable computers to gain a high-level understanding from digital images or videos [98]. The goal is to automate tasks that the human visual system can do, such as recognizing objects, identifying people, and detecting events. CV has become a cornerstone of modern AI, with applications ranging from self-driving cars and medical imaging to facial recognition and augmented reality. DL, in particular, has revolutionized CV, with convolutional neural networks (CNNs) and, more recently, vision transformers and state space models achieving remarkable results on a wide array of visual tasks including:

- **Image classification**: assigning a single category label to an entire image.
- **Object detection**: locating and classifying multiple objects within an image using bounding boxes.
- **Semantic segmentation**: assigning a class label to every pixel in an image.
- **Instance segmentation**: identifying and delineating each distinct object instance in an image at the pixel level, even for objects of the same class.
- **Object tracking**: following the trajectory of one or more objects over a sequence of video frames.

3.1 Domain generalization

3.1.1 Problem formulation

In the domain generalization (DG) problem, we consider a scenario where training and test data are drawn from different but related distributions [1, 13]. Visually, this is perceived as stylistic variations between samples, *e.g.*, the visual differences between a photo and a sketch of a cat (see also Figure 3.1). Let $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_N\}$ denote a set of N source domains, where each domain $\mathcal{D}_i$ is characterized by a joint distribution $P_i(X, Y)$ over input-output pairs (x, y). Here, $x \in \mathcal{X}$ represents the input space (*e.g.*, images) and $y \in \mathcal{Y}$ represents the label space (*e.g.*, class labels).

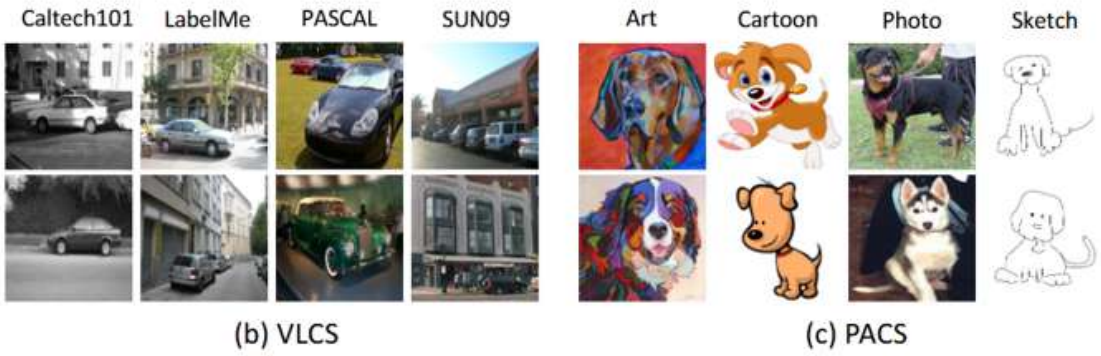


Figure 3.1. *Examples of domain shifts in two common domain generalization benchmarks. Source: [1]*

During training, we have access to labeled data from multiple source domains:

$$\mathcal{S} = \bigcup_{i=1}^{N} \mathcal{S}_i = \bigcup_{i=1}^{N} \{(x_j^{(i)}, y_j^{(i)})\}_{j=1}^{n_i}, \tag{3.1}$$

where $\mathcal{S}_i = \{(x_j^{(i)}, y_j^{(i)})\}_{j=1}^{n_i}$ denotes the training set from domain $\mathcal{D}_i$ with n_i samples, and $(x_j^{(i)}, y_j^{(i)}) \sim P_i(X, Y)$.

The goal of DG is to learn a predictive function $f : \mathcal{X} \to \mathcal{Y}$ that generalizes well to an unseen target domain $\mathcal{D}_t$ with distribution $P_t(X, Y)$, where $\mathcal{D}_t \notin \mathcal{D}$. Formally, we aim to minimize the expected risk on the target domain:

$$R_t(f) = \mathbb{E}_{(x,y)\sim P_t(X,Y)}[\ell(f(x), y)], \tag{3.2}$$

where $\ell(\cdot, \cdot)$ is a loss function (*e.g.*, cross-entropy for classification).

Since we have **no access to target domain data during training**, we minimize the empirical risk across all source domains:

$$\hat{R}_{\text{emp}}(f) = \frac{1}{|\mathcal{S}|} \sum_{i=1}^{N} \sum_{j=1}^{n_i} \ell(f(x_j^{(i)}), y_j^{(i)}). \tag{3.3}$$

However, minimizing only the empirical risk may lead to overfitting to source domains.

Therefore, domain generalization methods typically optimize a regularized objective:

$$\min_f \hat{R}_{\text{emp}}(f) + \lambda\Omega(f), \tag{3.4}$$

where $\Omega(f)$ is a regularization term that encourages domain-invariant representations, and $\lambda > 0$ is a regularization parameter.

3.1.2 Disambiguation from related problems

DG is often conflated with several related problems. In order to provide a clear definition, it is essential to highlight the key distinctions from these other paradigms:

Domain Adaptation (DA)

The core difference between DG and DA lies in the availability of target domain data during the training phase. In DA, the model has access to unlabeled or limited labeled data from a set of specific, known target domains [99, 100, 101]. The objective is to leverage this data to align the source and target distributions. On the contrary, DG operates under a strict constraint: it assumes no access whatsoever to any data from the eventual target domain during training.

Test-Time Adaptation (TTA)

TTA is distinguished from DG by the timing of model modification. In TTA, the model is permitted to adapt its parameters or predictions using the incoming target domain data at the time of inference [102, 103]. This typically involves leveraging self-supervision or techniques like entropy minimization on the test batch itself to improve performance. The foundational premise of pure DG, however, is that the model trained on the source domains must remain fixed and immutable at test time, making predictions without any subsequent updates or adaptation procedures.

Few/Zero-Shot Learning (FSL/ZSL)

Both FSL [104, 105] and ZSL [106, 107] address generalization challenges primarily focused on the class space, rather than the domain shift. FSL is concerned with the ability to classify new classes with only a handful of examples (*e.g.*, one to five per class), emphasizing rapid learning from minimal data. ZSL goes a step further, aiming to classify instances of classes never seen training by utilizing auxiliary semantic information, such as attributes or textual descriptions. In contrast, DG assumes a consistent class space between the source and target environments, and its difficulty stems solely from the distributional shift across domains.

Transfer Learning (TL)

Finally, TL represents the broadest conceptual umbrella, defining the entire paradigm of applying knowledge from one task or domain to another. Consequently, DG is not

an alternative to TL, but rather a specific, highly constrained instance of it [108, 109, 110]. DG differentiates itself within this field by tackling the challenging scenario where knowledge must be transferred to a target domain that is completely inaccessible and unknown throughout the training process.

The comparison between the aforementioned problems is better illustrated in Tab. 3.1. We must note that, in the majority of DG applications, the target label space is considered identical to the source one. This assumption is also made in the setting of our approach.

	N		$\mathbb{P}_{XY}^{S}$? $\mathbb{P}_{XY}^{T}$		$\mathcal{Y}_S$? $\mathcal{Y}_T$		Access to $\mathbb{P}_{X}^{T}$?
	$=1$	>1	$=$	$\neq$	$=$	$\neq$	
Domain Adaptation	✓	✓		✓	✓	✓	✓
Test-Time Adaptation	✓	✓		✓		✓	✓†
Few/Zero-Shot Learning	✓			✓		✓	
Transfer Learning	✓	✓		✓	✓	✓	
Domain Generalization	✓	✓		✓	✓	✓	

N: number of source domains/tasks. $\mathbb{P}_{XY}^{S/T}$: source/target joint distribution. $\mathcal{Y}_{S/T}$: source/target label space. $\mathbb{P}_{X}^{T}$: target marginal. †: Limited in quantities, like a single example or mini-batch.

Table 3.1. *Comparison between domain generalization and its related problems [1].*

3.1.3 Domain generalization in the era of pretraining

The classical paradigm of DG is built upon the core assumption of a truly ***unseen* target domain**, where a model must generalize its knowledge from a limited set of source domains. However, web-scale pretraining for foundation models, particularly vision-language models like CLIP [49], compels a re-evaluation of the standard DG paradigm. These models are trained on vast, unfiltered datasets like LAION [111], which contain billions of image-text pairs scraped from the internet. Therefore, the data used in traditional DG benchmarks (*e.g.*, sketches, paintings, cartoons in the PACS dataset) are almost certainly represented, at least stylistically, if not identically, within these massive pretraining datasets. This leads to a critical issue of **domain** (or even **sample**) **contamination**, where the backbone's prior exposure to the target distribution is a strong possibility [14]. As a result, the model is not learning to generalize from scratch; rather, it uses its pre-existing knowledge to perform a downstream task in a new context.

This has led researchers to argue that for foundation models the problem would be more accurately described as *zero-shot domain transfer*, implying that the model's performance stems from its vast prior knowledge rather than the specific DG algorithm applied during fine-tuning. Therefore, the research community has already moved towards the creation of more challenging benchmarks with carefully curated samples to accurately assess generalization capabilities [14].

3.1.4 Related work

The field of DG has seen a rapid evolution of techniques aimed at learning models that are robust to unseen domain shifts. The core challenge lies in learning representations that are invariant to domain-specific variations while preserving essential semantic information. Early approaches laid a foundational groundwork, which has since been expanded upon by more sophisticated methods that can be broadly categorized into several directions [1].

Domain alignment

A common approach in DG is to explicitly align the feature distributions from multiple source domains. The goal is to learn a domain-agnostic feature space where a single classifier can perform well regardless of the input domain. This is often achieved through adversarial learning [15, 16, 17] or by minimizing statistical distance metrics (*e.g.*, ℓ_2, Wasserstein [112] and Kullback-Leibler divergence [113]) [18, 19]. The seminal work on Domain-Adversarial Neural Networks (DANNs) [114], while designed for domain adaptation, pioneered the use of a domain discriminator with a gradient reversal layer to enforce feature invariance. Building on similar principles, methods like **SagNet** (Style-Agnostic Network) [115] mitigate domain-specific information by learning to ignore features that are highly correlated with the domain, effectively aligning the representations by focusing on style-agnostic content.

Data augmentation

Instead of aligning existing domains, another strategy is to synthesize novel domains to increase the diversity of the training data. Augmentation can take place in image space [20, 21] (by applying various transformations, *e.g.*, random crop, rotation, horizontal/vertical flip) or, more often, in the more compact feature space [23], for greater efficiency. A prominent example of the latter is **MixStyle** [22], which probabilistically mixes the feature statistics (mean and standard deviation) between instances from different source domains.

Learning strategies

This category includes a range of methods that focus on the optimization process to find a solution that generalizes better.

- **Worst-case optimization:** These methods are inspired by robust optimization principles. **GroupDRO** [24] minimizes the loss on the worst-performing domain, while **VReX** [25] regularizes the model by penalizing the variance of the losses across domains, both aiming for more uniform performance.
- **Self-challenging and regularization: RSC** (Representation Self-Challenging) [26] proposes a simple yet powerful regularization technique where the model is prevented from relying on dominant features by dropping activations with the highest gradients during training, forcing it to learn from a more diverse set of cues. Similarly, **ARM** (Adaptive Risk Minimization) [27] uses a meta-learning approach to find a

parameterization that is robust to worst-case empirical risk over a set of plausible data distributions.

- **Advanced optimization:** Other works have explored the geometry of the loss landscape. **SWAD** (Stochastic Weight Averaging Densely) [28] finds a wider, flatter minimum by averaging model weights over training iterations, a technique shown to improve generalization. **SAGM** (Sharpness-Aware Gradient Matching) [29] extends this by seeking a flat minimum where the gradients from different domains are well-aligned, combining sharpness-aware minimization with gradient consensus.

- **Feature space learning: PCL** (Proxy-based Contrastive Learning) [30] replaces the standard sample-to-sample contrastive learning with proxy-to-sample relations, directly addressing the positive alignment issue appearing in conventional contrastive learning applications.

Representation learning

Another research direction aims to learn causal mechanisms or disentangled representations, separating domain-specific (stylistic) features from domain-invariant (semantic) features. **MTL** (Marginal Transfer Learning) [31] reframes DG as a supervised learning problem where the original feature space is augmented with the marginal distribution of feature vectors to leverage information about the test task's feature distribution. **iDAG** (invariant Directed Acyclic Graph) [32] reframes DG as an invariant causal graph searching problem and enforces constraints to guarantee acyclicity and eliminate spurious correlations. It also employs a prototype-based dataset proxy technique to avoid traversing the whole dataset in every training step. Furthermore, iDAG incorporates hierarchical contrastive learning to align the latent causal factors and improve the representativeness of the prototypes. **GMDG** (General Multi-Domain Generalization) [33] relaxes constraints on static target marginal distribution and distills a general learning objective into four practical components: two terms for learning domain-independent conditional features and maximizing a posterior, plus two regularization terms that incorporate prior information and suppress invalid causality.

Architectural advancements

ViT-based. While the majority of DG research has been algorithm-centric and backbone-agnostic, a recent trend explores how architectural choices can inherently improve domain generalization. The shift from CNNs to Vision Transformers (ViTs) has opened new avenues. **SDViT** [34] was an early work that adapted ViTs for DG by proposing a style diversification module specifically for the transformer architecture. **GMoE** [35] leveraged a Mixture-of-Experts architecture, where different "experts" (sub-networks) can specialize in different data characteristics, which can be beneficial for handling diverse domains.

VMamba-based. The latest advancements have begun to explore State Space Models (SSMs) (see also 3.2) like Mamba [2] and its adaptation for vision tasks VMamba [9] for

their efficiency and ability to model long-range dependencies. **DGMamba** [36] is the first work to use the VMamba architecture in the context of DG. In particular, it attempts to enforce similar hidden states across domains and shift the model's attention towards the object. **DGFamba** [37] uses flow factorization to align the pre- and post- hallucinated state embedding in the latent space.

3.2 State space models

3.2.1 Motivation

For several years, the transformer architecture has been the standard for sequence modeling, achieving state-of-the-art results across many tasks [38]. Its core component, the self-attention mechanism, enables the model to capture complex, long-range dependencies by computing pairwise interactions between all tokens in a sequence. However, this comprehensive modeling capability comes with a significant computational burden: the time and memory complexity of self-attention scales quadratically with the sequence length L as $O(L^2)$. This **quadratic scaling** renders **transformers** prohibitively expensive for processing the ultra-long sequences found in high-resolution vision, genomics, and long-form text analysis.

This fundamental limitation has spurred research into more efficient architectures that can maintain strong performance while scaling linearly with sequence length. State Space Models (SSMs) have emerged as a particularly promising alternative. The **Mamba** architecture represents a breakthrough in this area, proposing a **selective SSM** that not only **scales linearly** ($O(L)$) but also demonstrates **performance competitive** with, and often superior to, state-of-the-art transformers [2].

This section provides a detailed analysis of the SSM fundamentals, traces their evolution into the Mamba model, and explores its adaptation for computer vision.

3.2.2 Preliminaries

Originating from classical control theory, SSMs are a class of models that map a 1D input signal $u(t)$ to an output signal $y(t)$ via an intermediate higher-dimensional latent state $x(t) \in \mathbb{R}^N$.

Continuous-time formulation

A linear, time-invariant (LTI) SSM is defined by a system of first-order ordinary differential equations (ODEs):

$$\frac{dx(t)}{dt} = \mathbf{A}x(t) + \mathbf{B}u(t) \tag{3.5}$$

$$y(t) = \mathbf{C}x(t) + \mathbf{D}u(t) \tag{3.6}$$

Here, $\mathbf{A} \in \mathbb{R}^{N\times N}$ is the state matrix, $\mathbf{B} \in \mathbb{R}^{N\times 1}$ is the input matrix, $\mathbf{C} \in \mathbb{R}^{1\times N}$ is the output matrix, and $\mathbf{D} \in \mathbb{R}^{1\times 1}$ is the feedthrough matrix. The parameter N denotes the dimensionality of the hidden state.

Discrete-time formulation

For use in digital systems, which dominate the practical applications of ML, the continuous-time model must be discretized. Using a sampling step size Δ, the continuous parameters $(\mathbf{A}, \mathbf{B})$ are transformed into discrete counterparts $(\bar{\mathbf{A}}, \bar{\mathbf{B}})$. A common

discretization method is the zero-order hold (ZOH), which yields the following discrete-time recurrence:

$$x_k = \bar{\mathbf{A}} x_{k-1} + \bar{\mathbf{B}} u_k \tag{3.7}$$

$$y_k = \mathbf{C} x_k + \mathbf{D} u_k \tag{3.8}$$

The discrete matrices are derived from the continuous ones as follows:

$$\bar{\mathbf{A}} = e^{\Delta \mathbf{A}}, \qquad \bar{\mathbf{B}} = (\Delta \mathbf{A})^{-1}(e^{\Delta \mathbf{A}} - \mathbf{I})\Delta \mathbf{B}. \tag{3.9}$$

This **recurrent** formulation is computationally efficient for autoregressive inference, scaling linearly with sequence length. However, it is not practical for training on modern GPU hardware due to its sequentiality. Alternatively, the model can be expressed as a GPU-friendly **global convolution**:

$$y = \bar{\mathbf{K}} * u, \tag{3.10}$$

where the convolutional kernel $\bar{\mathbf{K}} \in \mathbb{R}^L := (\mathbf{C}\bar{\mathbf{B}}, \mathbf{C}\bar{\mathbf{A}}\bar{\mathbf{B}}, \ldots, \mathbf{C}\bar{\mathbf{A}}^{L-1}\bar{\mathbf{B}})$ is derived from the SSM parameters. This non-circular convolution can be computed efficiently with Fast Fourier Transform (FFT), allowing for highly parallelizable training [7].

3.2.3 The evolution from S4 to Mamba

Structured State Space (S4) sequence models

The S4 [7], a pivotal development that made SSMs competitive with advanced recurrent and convolutional models, introduced two key ideas:

1. **HiPPO initialization:** The state matrix A is initialized using the **High-order Polynomial Projection Operator (HiPPO)** framework, a theoretically grounded method designed to enable the model to effectively compress and remember long-term dependencies from an input history [6]. The core idea behind HiPPO is to maintain an online, continuous representation of a function's history by projecting it onto a basis of orthogonal polynomials. Specifically, it uses a basis of shifted Legendre polynomials to approximate the entire history of the input signal $u(\tau)$ for $\tau \leq t$. The coefficients of this polynomial approximation form the latent state vector $x(t)$. The key insight of the HiPPO framework is that the evolution of these coefficients over time can be described by a simple linear ODE:

$$\frac{dx(t)}{dt} = \mathbf{A}x(t) + \mathbf{B}u(t)$$

This provides a principled way to derive the state matrix $\mathbf{A}$. For a state of size N,

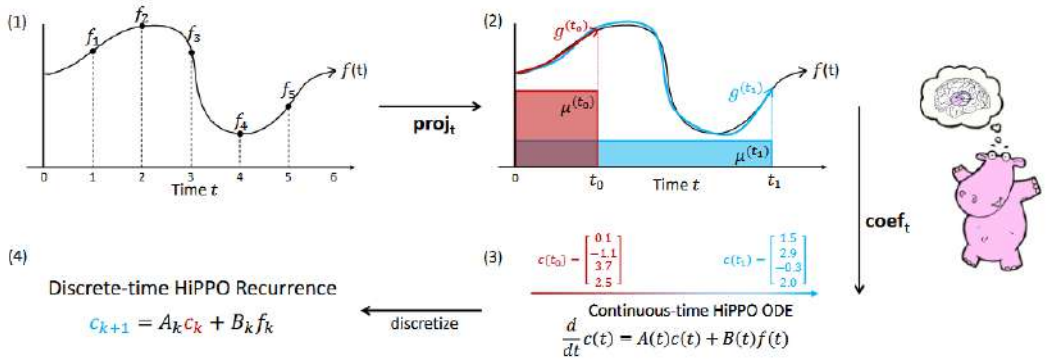


Figure 3.2. *Illustration of the HiPPO framework [6]. **(1)** For any function f, **(2)** at every time t there is an optimal projection $g(t)$ of f onto the space of polynomials, with respect to a measure $\mu(t)$ weighing the past. **(3)** For an appropriately chosen basis, the corresponding coefficients $c(t) \in \mathbb{R}^N$ representing a compression of the history of f satisfy linear dynamics. **(4)** Discretizing the dynamics yields an efficient closed-form recurrence for online compression of time series $(f_k)_{k\in\mathbb{N}}$.*

the HiPPO matrix is specifically defined as:

$$(\mathbf{A})_{nk} = \begin{cases} (2n+1)^{1/2}(2k+1)^{1/2}, & \text{if } n > k \\ -(n+1), & \text{if } n = k \\ 0, & \text{if } n < k \end{cases} \tag{3.11}$$

By initializing the SSM with this specific matrix structure, the model is inherently equipped to maintain a state that represents an optimal, low-dimensional polynomial approximation of the input's history.

2. **Matrix Structure:** A critical challenge in making SSMs practical is the computational cost associated with the $N \times N$ state matrix $\mathbf{A}$, where N can be large. Operations like the matrix exponential, $e^{\Delta\mathbf{A}}$, required for discretizing the system, have a cubic complexity of $O(N^3)$ for a general matrix, which is computationally prohibitive. To overcome this bottleneck, the S4 model imposes a specific algebraic structure on the $\mathbf{A}$ matrix, known as **Diagonal Plus Low-Rank (DPLR)**. As the name suggests, a DPLR matrix $\mathbf{A}$ is defined as:

$$\mathbf{A} = \mathbf{\Lambda} - \mathbf{P}\mathbf{Q}^* \tag{3.12}$$

where $\mathbf{\Lambda} \in \mathbb{C}^{N\times N}$ is a diagonal matrix, and $\mathbf{P}, \mathbf{Q} \in \mathbb{C}^{N\times R}$ with the rank R being a small integer (*e.g.*, 1 or 2), making $R \ll N$.

This DPLR structure is not arbitrary; the computation of the spectrum of $\bar{\mathbf{K}}$, the reduction of exponentiation to matrix inverse, as well as the equivalence of the diagonal case to the highly optimized computation of a Cauchy kernel allow for resolving the bottleneck observed in previous SSMs [7].

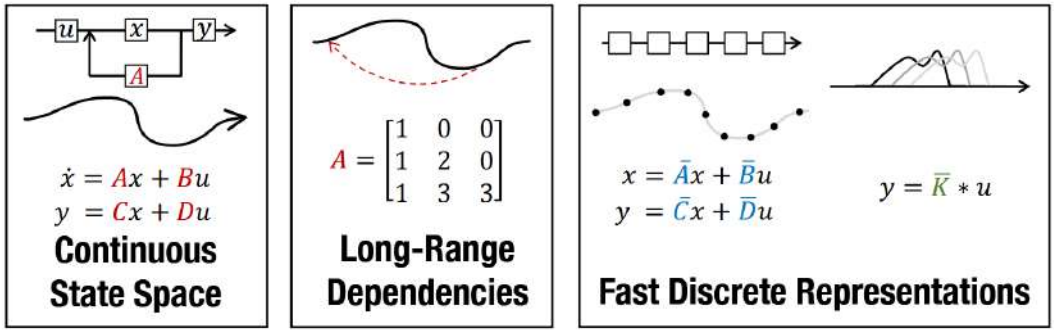


Figure 3.3. ***(Left)*** *State Space Models (SSM) parameterized by matrices* $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ *map an input signal* $u(t)$ *to output* $y(t)$ *through a latent state* $x(t)$. ***(Center)*** *Recent theory on continuous-time memorization derives special* A *matrices that allow SSMs to capture long-range dependencies mathematically and empirically.* ***(Right)*** *SSMs can be computed either as a recurrence (left) or convolution (right). However, materializing these conceptual views requires utilizing different representations of its parameters (red, blue, green) which are very expensive to compute. S4 introduces a parameterization that efficiently swaps between these representations, allowing it to handle a wide range of tasks, be efficient at both training and inference, and excel at long sequences. Source: [7]*

The aforementioned ideas enabled S4 to retain the theoretical benefits of a large latent state N for modeling long-range dependencies while ensuring that the model remains computationally efficient for practical use.

Despite its breakthroughs, S4 continues to be **time-invariant** and **input-agnostic**. The system matrices $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \Delta)$ are fixed after training and do not adapt to the input sequence, limiting the model's capacity for content-aware reasoning.

Selective State Space (S6) models

The core innovation of Mamba is the introduction of a selection mechanism, which is realized by making the SSM parameters **input-dependent** [2]. This transforms the system from a time-invariant one to a time-varying one, allowing the model to dynamically modulate its behavior based on the content of the input. The model is referred to as S6, because it is essentially a S4 model equipped with a *selection* mechanism and utilizing a parallel *scan* algorithm.

Specifically, the parameters Δ, $\mathbf{B}$, and $\mathbf{C}$ are no longer fixed but are instead generated from the input sequence u_t:

$$\mathbf{B}_t = s_B(x_t) \tag{3.13}$$

$$\mathbf{C}_t = s_C(x_t) \tag{3.14}$$

$$\Delta_t = \tau_\Delta(\text{Parameter} + s_\Delta(x_t)) \tag{3.15}$$

where s_B, s_C, s_Δ are learnable projection functions (typically linear layers), and τ_Δ (*e.g.*, softplus) ensures $\Delta_t > 0$. The input-dependent step size Δ_t is particularly crucial for selectivity. A large Δ_t causes the state x_t to focus on the current input u_t and effectively

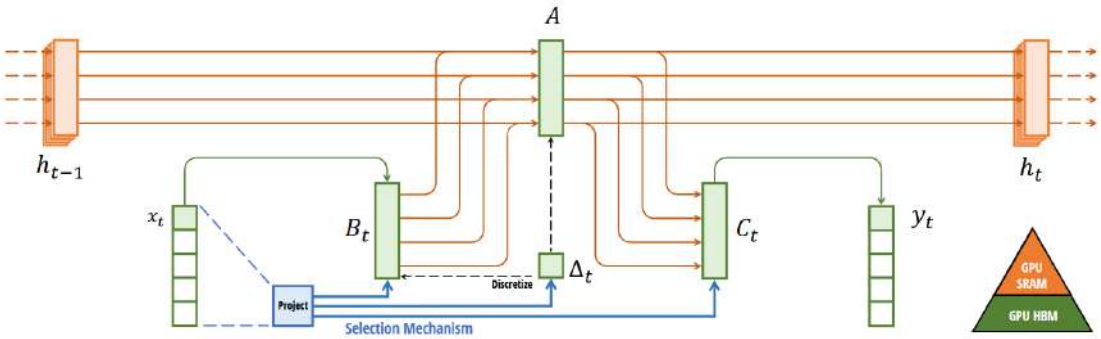


Figure 3.4. *Overview of the Mamba architecture. Mamba's selection mechanism combines input-dependent dynamics with a careful hardware-aware algorithm to only materialize the expanded states in more efficient levels of the GPU memory hierarchy, achieving significant speedup. Source: [2]*

reset or forget its previous history. Conversely, a small Δ_t allows the state to persist and remember information from the past. This gives the model a fine-grained mechanism to decide which information to propagate and which to filter out.

The Mamba architecture

The Mamba architecture integrates the aforementioned selective S6 scan into a simplified block structure, reminiscent of a transformer block but without the quadratic self-attention. However, a key challenge with time-varying SSMs is that they cannot be materialized as a single global convolution, making parallel training difficult. Mamba overcomes this with a **hardware-aware parallel scan algorithm** [2]. The efficiency of the algorithm stems from the application of three techniques:

1. The algorithm performs (computational) **kernel fusion** in order to decrease the number of memory IOs, achieving significant speedup.
2. The SSM parameters are loaded directly from slow HBM (High-Bandwidth Memory) to **fast SRAM** (Static Random-Access Memory), where discretization and recurrence are also performed. Afterwards, the final outputs are written back to HBM.
3. The **intermediate states** are not stored but **recomputed** in the backward pass when the inputs are loaded from HBM to SRAM.

Consequently, the fused selective scan layer has the same memory footprint as an optimized transformer implementation with FlashAttention [116].

3.2.4 Mamba in vision

Adapting the inherently 1D Mamba model to 2D visual data presents a unique set of challenges. The standard approach involves partitioning an image into a grid of patches and flattening them into a 1D sequence. However, this process disrupts the 2D spatial locality.

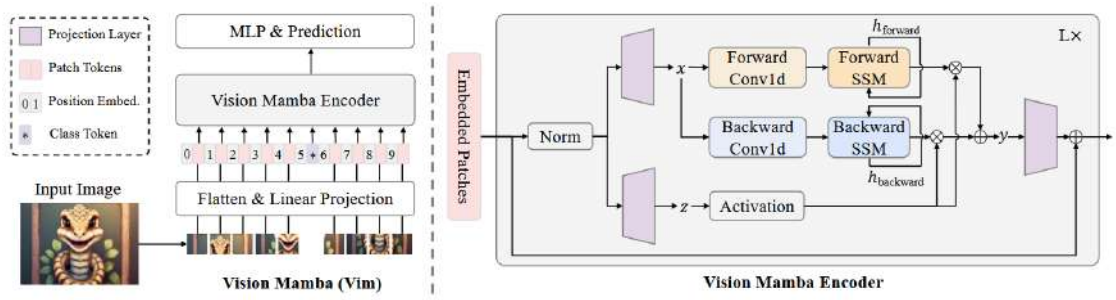


Figure 3.5. *Overview of Vim. The input image is initially split into patches and projected into patch tokens. Finally, the sequence of tokens is sent to the Vim encoder. To perform classification, an extra learnable classification token is concatenated to the patch token sequence. Contrary to Mamba for text sequence modeling, Vim encoder processes the token sequence bidirectionally. Source: [8]*

Vision Mamba (Vim)

Patch embedding. Vim represents one of the first successful adaptations of the 1D Mamba architecture for 2D computer vision tasks [8]. The process begins by adopting the standard approach from Vision Transformers (ViT): the input image is first partitioned into a grid of non-overlapping 2D patches and then flattened into a 1D sequence (patch embedding), to which a class token is prepended. A learnable positional embedding is added to this new sequence, in order to retain spatial information that is lost during flattening, and the resulting sequence becomes the input to the stack of Vim blocks.

The Vim block. The core of the Vim model is the **Vim block**, which replaces the multi-head self-attention mechanism of a standard Transformer block with a bidirectional Mamba module. A Vim block processes an input sequence Z_{l-1} to produce an output Z_l through two main sub-layers:

1. **Bidirectional Mamba module:** This module is responsible for spatial context aggregation. The input sequence Z_{l-1} is first normalized using LayerNorm. It is then processed in parallel by two Mamba blocks: one that scans the sequence in the forward direction and one that scans a reversed version of the sequence. The mathematical flow is as follows:

$$Z'_l = \text{LayerNorm}(Z_{l-1}) \tag{3.16}$$
$$Y_{\text{fwd}} = \text{Mamba}(Z'_l) \tag{3.17}$$
$$Y_{\text{bwd}} = \text{rev}(\text{Mamba}(\text{rev}(Z'_l))) \tag{3.18}$$
$$Y_l = Z_{l-1} + Y_{\text{fwd}} + Y_{\text{bwd}} \tag{3.19}$$

   The outputs from both directions are added together and then combined with the original input through a residual connection (Equation 3.19). This bidirectional scan allows every patch to efficiently gather information from all other patches, effectively creating a global receptive field.

2. **MLP sub-layer:** The output from the Mamba module, Y_l, is then passed to a standard feed-forward network, which provides per-token feature transformation. This sub-layer consists of a LayerNorm, a two-layer Multi-Layer Perceptron (MLP) with a GELU activation function, and a second residual connection:

$$Y_l' = \text{LayerNorm}(Y_l) \tag{3.20}$$

$$Z_l = Y_l + \text{MLP}(Y_l') \tag{3.21}$$

By stacking these Vim blocks, the model can build a hierarchical representation of the image, making it a powerful and computationally efficient backbone for various vision tasks.

VMamba

While initial adaptations like Vim demonstrated the potential of Mamba for vision, the VMamba model was a significant leap forward, designed to establish State Space Models (SSMs) as a general-purpose, high-performance backbone for computer vision, on par with leading Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) [9]. VMamba's core contribution is a novel block design that effectively addresses the challenge of applying a 1D sequence model to 2D spatial data.

The Cross-Scan Module (CSM). The central problem with applying SSMs to images is that flattening the 2D grid of patches into a 1D sequence breaks spatial locality. A simple bidirectional scan can only partially recover this lost information. To solve this, VMamba introduces the **Cross-Scan Module (CSM)**, a sophisticated mechanism that enables information to flow across the image in multiple directions.

Instead of a simple forward and backward pass, the CSM processes the 2D feature map by unrolling it into sequences along four different directions:

1. Top-left to bottom-right
2. Bottom-right to top-left
3. Top-right to bottom-left
4. Bottom-left to top-right

This is achieved by strategically rearranging the flattened sequence of tokens before feeding them into four parallel S6 blocks (the core selective scan from Mamba). After processing, the four output sequences are merged back into their original 2D arrangement. This multi-directional scan ensures that every pixel (or patch token) can aggregate contextual information from its entire spatial neighborhood efficiently and effectively, equipping the model with an extended, 2D-aware receptive field.

The Visual State Space (VSS) block. The VSS block is the main building block of the VMamba architecture. It is designed to be a powerful and efficient replacement for the self-attention block in a Vision Transformer.

The data flow through a VSS block is as follows:

1. The input tensor $X_{in} \in \mathbb{R}^{H \times W \times C}$ from the previous layer is first processed by a linear layer to expand its channel dimension.
2. The result is then passed through the CSM as described above, which performs the four-directional spatial mixing.
3. The output of the CSM is passed through a non-linearity (*e.g.*, SiLU) and another linear layer that projects it back to the original channel dimension.
4. A residual connection adds the original input X_{in} to the output of the block, ensuring stable training.

The overall structure can be summarized as:

$$X_{out} = X_{in} + \text{Linear}(\text{CSM}(\text{Linear}(X_{in}))) \tag{3.22}$$

By stacking these VSS blocks, VMamba creates a hierarchical backbone that can learn rich feature representations at multiple scales. This design allows VMamba to achieve state-of-the-art performance on a wide range of dense prediction tasks like object detection and semantic segmentation, demonstrating that its 2D-aware SSM approach is both powerful and computationally efficient.

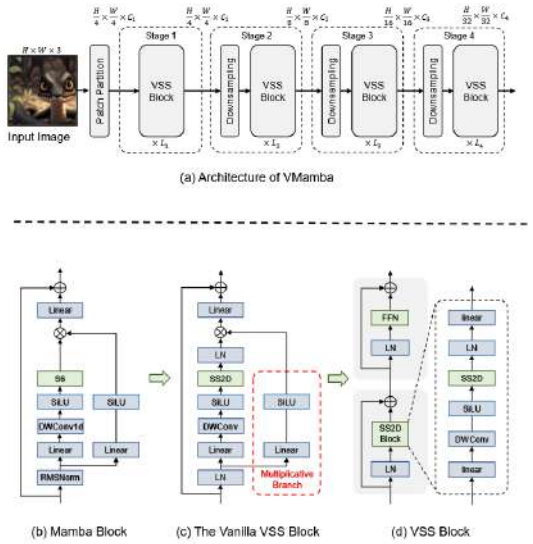


Figure 3.6. *(**a**) The input image is first partitioned into patches by a stem module, resulting in a 2D feature map. Without incorporating additional positional embeddings, multiple network stages are employed to create hierarchical representations. Specifically, each stage comprises a down-sampling layer (except for the first stage), followed by a stack of VSS blocks. (**b**)-(**d**) The vanilla VSS block is formulated by replacing the S6 module with the newly proposed 2D-Selective-Scan (SS2D) module. The improved VSS block consists of a single network branch with two residual modules, mimicking the architecture of a vanilla transformer block. Source: [9]*

2D-Selective-Scan (SS2D). Data forwarding in SS2D consists of three steps:

1. **Cross-scan**: SS2D first unfolds the input patches into sequences along four distinct traversal paths.

2. **Selective scanning**: Each patch sequence is then processed in parallel using a separate S6 block.

3. **Cross-merge**: The resultant sequences are reshaped and merged to form the context-aware output map.

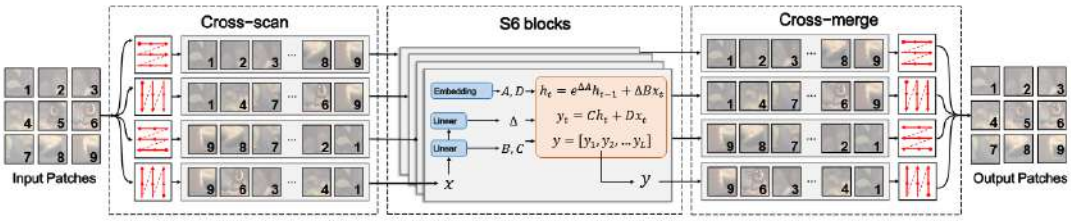


Figure 3.7. *Illustration of SS2D. Source: [9]*

Variants and extensions

Architectures for dense prediction tasks. For tasks like segmentation and object detection that require fine-grained spatial understanding, several specialized architectures have been proposed. **SegMamba** [39] employs a hierarchical Mamba backbone to extract multi-scale features, similar to architectures like the Swin Transformer. It then uses a lightweight Mamba-based decoder to effectively fuse these features for precise pixel-level semantic segmentation. Other variants like **U-Mamba** [40] and **Mamba-UNet** [41] integrate Mamba blocks into the classic U-Net encoder-decoder architecture, a dominant model in medical image segmentation. Mamba blocks, equipped with 2D-aware scanning, often replace the standard convolutional layers in the encoder's bottleneck or decoder's expansion path. This combination of the proven multi-scale feature fusion of U-Net with the superior long-range dependency modeling of Mamba yields remarkable results in medical imaging.

Hybrid architectures. Recognizing the complementary strengths of different architectures, hybrid models have emerged to leverage the best of both worlds. **ConvMamba** [42], in particular, follows a hybrid design where the early stages of the network are composed of efficient convolutional layers. These layers excel at capturing local, low-level features like edges and textures with a strong inductive bias. The deeper stages of the network then transition to 2D-aware Mamba blocks, which are better suited for modeling the global, long-range relationships between the extracted local features. This approach combines the efficiency and locality of CNNs with the global context modeling of SSMs. Furthermore, **MambaVision** [43] introduces an improved, vision-friendly Mamba block and incorporates self-attention blocks at the final stages of processing to enhance the model's ability to capture context and long-range spatial relationships.

3.3 Contrastive learning

Contrastive learning (CL) has emerged as a powerful paradigm for learning representations in an embedding space, where similar (positive) instances are mapped close together while dissimilar (negative) instances are pushed apart. Similarity can be qualitative or quantitative, and in most applications refers to semantic alignment.

3.3.1 Mathematical framework

The CL framework operates on pairs of data points. Given a dataset $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$, we define positive pairs (x_i, x_j^+) as similar instances and negative pairs (x_i, x_j^-) as dissimilar instances. An encoder network $f_\theta : \mathcal{X} \to \mathbb{R}^d$ maps inputs to a d-dimensional embedding space, often followed by a projection head $g_\phi : \mathbb{R}^d \to \mathbb{R}^{d'}$, where $\mathbb{R}^{d'}$ is typically a lower-dimensional space where the contrastive loss is applied. The projection head typically produces normalized embeddings $z = g_\phi(f_\theta(x))$ with $\|z\| = 1$.

The core of CL is the InfoNCE loss [44], which maximizes mutual information between positive pairs:

$$\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E}_{(x,x^+)}\left[\log \frac{\exp(\text{sim}(z, z^+)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(z, z_j)/\tau)}\right], \tag{3.23}$$

where $z = g_\phi(f_\theta(x))$, as stated above, $\text{sim}(z_i, z_j) = z_i^\top z_j$ is the **cosine similarity**, τ is a temperature parameter controlling the concentration of the distribution, and N is the total number of samples in the batch (including one positive and $N-1$ negatives).

The temperature parameter τ plays a critical role in the optimization dynamics. Lower temperatures create sharper distributions that emphasize hard negatives, while higher temperatures produce smoother distributions. Empirically, $\tau \in [0.07, 0.5]$ has been found effective across different tasks [45].

3.3.2 Variants and extensions

The need for (many) negatives

SimCLR [45] applies a composition of random transformations (crop, color jitter, blur) to create two augmented views of the same image: $(x_i, x_i^+) = (\text{aug}_1(x), \text{aug}_2(x))$. The framework employs large batch sizes (*e.g.*, 4096) to provide sufficient negative samples within each training step.

MoCo [46] addresses the batch size limitation through a *momentum encoder* and *memory bank*. The momentum encoder f_{θ_m} is updated via exponential moving average: $\theta_m \leftarrow m\theta_m + (1-m)\theta$, where $m \in [0.9, 0.999]$ is the momentum coefficient. This creates a large, consistent dictionary of negative samples without requiring massive batches:

$$\mathcal{L}_{\text{MoCo}} = -\log \frac{\exp(q^T k^+/\tau)}{\exp(q^T k^+/\tau) + \sum_{i=1}^{K} \exp(q^T k_i/\tau)} \tag{3.24}$$

where $q = f_\theta(x)$ is the query, $k^+ = f_{\theta_m}(x^+)$ is the positive key, and $\{k_i\}$ are negative keys from the memory bank.

Learning without explicit negatives

BYOL (Bootstrap Your Own Latent) [47] showed that explicit negative samples might not be necessary. It uses two networks: an *online* network and a *target network* (which is a momentum-updated version of the online one). The goal is to make the online network's prediction of a view match the target network's representation of another view. It prevents collapse through the momentum encoder.

SimSiam [48] simplified BYOL even further by showing that the momentum encoder is not necessary. It demonstrated that a simple *stop-gradient* operation was sufficient to prevent the model from collapsing to a trivial solution.

Cross-modal contrastive learning

CLIP (Contrastive Language-Image Pretraining) [49] represents a landmark achievement in cross-modal learning, fundamentally changing the landscape of vision models. Instead of training on a curated dataset with fixed labels, CLIP is pretrained on a massive, noisy dataset of 400 million image-text pairs scraped from the internet. CLIP trains image and text encoders jointly using a symmetric cross-modal objective. Given a batch of N image-text pairs $\{(I_i, T_i)\}_{i=1}^{N}$, CLIP processes it using two separate encoders: a Vision Transformer (ViT) for images and a text transformer for the corresponding captions. After computing the image embeddings $\mathbf{v}_i = f_v(I_i)$ and text embeddings $\mathbf{t}_i = f_t(T_i)$, CLIP optimizes:

$$\mathcal{L}_{\text{i2t}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\mathbf{v}_i^\top \mathbf{t}_i/\tau)}{\sum_{j=1}^{N}\exp(\mathbf{v}_i^\top \mathbf{t}_j/\tau)}, \tag{3.25}$$

$$\mathcal{L}_{\text{t2i}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\mathbf{t}_i^\top \mathbf{v}_i/\tau)}{\sum_{j=1}^{N}\exp(\mathbf{t}_i^\top \mathbf{v}_j/\tau)}, \tag{3.26}$$

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2}(\mathcal{L}_{\text{i2t}} + \mathcal{L}_{\text{t2i}}). \tag{3.27}$$

This symmetric formulation ensures that both image-to-text and text-to-image retrieval are jointly optimized. This has enabled CLIP's most significant emergent capability, *i.e.*, zero-shot classification. By creating text prompts like "*a photo of a* `class`," CLIP can classify images into arbitrary categories without ever being explicitly trained on them. It simply finds which text prompt's embedding is closest to the image's embedding in the shared space. This has established CLIP as a backbone for a wide range of vision tasks.

3.3.3 Modality gap

Empirical observation

Despite the success of methods like CLIP, the cross-modal CL optimization process has a well-documented side effect: it creates a highly anisotropic embedding space. Instead of embeddings being uniformly distributed across the entire feature hypersphere, they are pushed into a narrow, cone-like region [3]. This happens because the InfoNCE loss is

optimized by cosine similarity, which only cares about the angle between vectors, not their magnitude or absolute position. Consequently, the model learns that the most efficient way to separate a positive pair from thousands of negatives is to cluster all embeddings into a specific area of the space, maximizing angular separation within that cone. This inherent property of the training objective is a direct precursor to the modality gap [50].

Visually, if we were to plot the embeddings from both modalities, we would not see a single, mixed cloud of points. Instead, we would see two distinct clusters—one for images and one for text—separated by a noticeable gap, called the **modality gap**. This creates a disjointed topology where the two modalities occupy separate sub-regions of the shared space.

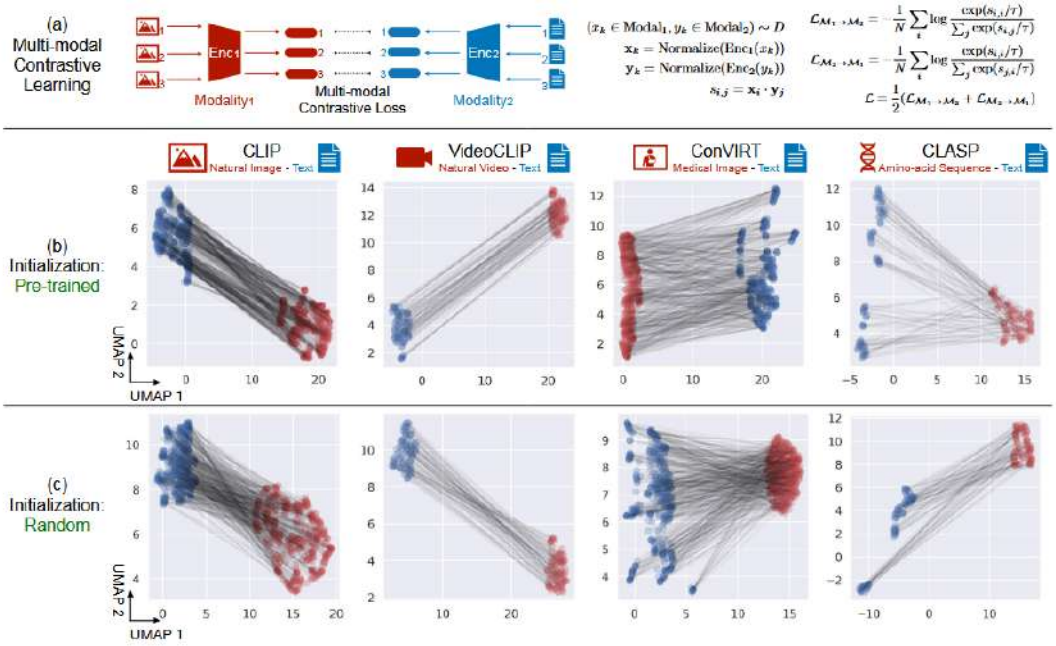


Figure 3.8. *The pervasive modality gap in multi-modal contrastive representation learning. (a) Overview of multi-modal contrastive learning. Paired inputs from two modalities (e.g., image-caption) are sampled from the dataset and embedded into the hypersphere using two different encoders. The loss function is to maximize the cosine similarity between matched pairs given all the pairs within the same batch. (b) UMAP visualization of generated embeddings from pre-trained models. Paired inputs are fed into the pre-trained models and the embeddings are visualized in 2D using UMAP (lines indicate pairs). We observe a clear modality gap for various models trained on different modalities. (c) UMAP visualization of generated embeddings from same architectures with random weights. Modality gap exists in the initialization stage without any training. Source: [3]*

Mathematical formulation

Given a set of N ℓ_2-normalized image embeddings $\mathcal{V} = \{\mathbf{v}_i\}_{i=1}^N$ and a corresponding set of ℓ_2-normalized text embeddings $\mathcal{T} = \{\mathbf{t}_i\}_{i=1}^N$, the respective centroids are calculated as their means:

$$\boldsymbol{\mu}_{\text{img}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i, \qquad \boldsymbol{\mu}_{\text{txt}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{t}_i. \tag{3.28}$$

According to [3], the modality gap can be formally quantified as the ℓ_2-norm of the difference between these two centroids:

$$\|\boldsymbol{\Delta}_{gap}\|_2 = \|\boldsymbol{\mu}_{\text{img}} - \boldsymbol{\mu}_{\text{txt}}\|_2. \tag{3.29}$$

Origins and causes

Architectural asymmetry. Vision-language models employ different encoder architectures for images (ResNet, ViT) and text (transformer), leading to distinct computational paths that create embeddings with different geometric properties even when trained jointly [3, 51].

Training dynamics. The contrastive objective optimizes relative similarities rather than absolute alignment. The symmetric InfoNCE loss:

$$\mathcal{L} = \frac{1}{2}\left(\mathcal{L}_{\text{i2t}} + \mathcal{L}_{\text{t2i}}\right)$$

encourages matching pairs to be more similar than non-matching pairs but does not explicitly enforce $\mathbf{v}_i = \mathbf{t}_i$. The temperature parameter τ creates a trade-off between alignment precision and separation of negatives [45].

Implications and consequences

Zero-shot classification. Tasks like zero-shot classification rely on cosine similarity, which only measures the angle between vectors, not their absolute positions. Since the gap is a systematic shift, it doesn't change the relative angles. The closest text embedding to an image embedding, in the sense of cosine similarity, will still be the correct one, leaving retrieval and zero-shot tasks unaffected [3].

Modality-specific information. Images and text may encode complementary information that cannot be perfectly aligned. Visual features capture spatial, compositional, and perceptual details, while text provides abstract, symbolic, and linguistic structure (information asymmetry). For example, the text description "*a photo at the beach that I took last summer*" contains temporal context ("*last summer*") which is visually imperceptible. Forcing a perfect, one-to-one mapping between the image and text embeddings could risk losing such modality-specific information. In this light, the modality gap might be a feature, rather than a bug. Images and text contain fundamentally different types of information and the gap plays an essential role in keeping them distinct within the joint embedding space [3].

Cross-modal generation. Text-guided image generation or editing typically tries to bridge the gap in order to effectively transfer semantic information from textual descriptions to visual outputs. Direct interpolation in the embedding space may not preserve semantic meaning [52].

Bridging the gap

Post-hoc centering. This strategy involves calculating the mean embedding for each modality from a large sample of data and then subtracting this mean from any new embeddings at inference time. This effectively moves the centroids of both distributions to the origin of the feature space, reducing the gap [3].

Parameter space sharing. AlignCLIP [51] proposes sharing the learnable parameters of the vision and language encoders and introduces a semantically-regularized intra-modality separation module, improving zero-shot classification, robustness to distribution shifts, and classification with linear probing.

Learned prior for translation. This approach addresses the modality gap not by altering the embedding space itself, but by explicitly learning a dedicated mapping function to translate representations from one modality's distribution to the other. A typical example of this strategy is DALL-E 2 [52]. Its generative process is decoupled into two stages: first, a *prior network* takes a CLIP text embedding as input and learns to generate a corresponding "ideal" CLIP image embedding. This prior's entire purpose is to bridge the modality gap by mapping a point in the "text region" of the space to its correct counterpart in the "image region." Only after this translation is the resulting image embedding used as a conditioning signal for the second stage, a *diffusion decoder*, which generates the final photorealistic image.

Temperature scheduling. Some works have explored linear and exponential schedules, as well as large values for the temperature variable at the later stages of training, so as to ensure uniformity on the unit hypersphere [53].

3.4 Generative modeling

Generative modeling is a fundamental paradigm in machine learning that aims to learn the underlying true probability distribution of observed data, enabling the generation of synthetic samples that resemble the true distribution. Given a dataset $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$ drawn from an unknown distribution $p_{\text{data}}(x)$, the objective of generative modeling is to approximate this distribution with a parameterized model $p_\theta(x)$ such that samples drawn from $p_\theta(x)$ are indistinguishable from those in the original dataset. This fundamental problem has driven decades of research, from early approaches based on Gaussian mixture models (GMMs), which approximate the true distribution via a superposition of Gaussian distributions, and hidden Markov models (HMMs) to modern deep learning frameworks that can generate high-resolution images, realistic text, and complex structured data, in general. The mathematical foundation of generative modeling rests on density estimation, where the goal is to minimize the divergence between the true and learned distributions, typically formulated as maximizing the log-likelihood $\mathcal{L}(\theta) = \mathbb{E}_{x\sim p_{\text{data}}}[\log p_\theta(x)]$ or equivalently minimizing the Kullback-Leibler divergence $D_{KL}(p_{\text{data}} \| p_\theta) = \mathbb{E}_{x\sim p_{\text{data}}}[\log p_{\text{data}}(x) - \log p_\theta(x)]$ [54, 55].

In the rest of the section, we will outline the most widely-used generative frameworks, including flow matching, which is used in our proposed method in chapter 4.

3.4.1 Variational autoencoders

Variational Auto-Encoders (VAEs) [56, 57, 4] represent a principled approach to generative modeling that combines variational inference with deep learning to learn latent variable models. The key insight underlying VAEs is the introduction of a latent variable that captures the underlying factors of variation in the data, enabling both efficient inference and generation through a tractable optimization framework.

Theoretical foundations

The VAE framework assumes that observed data x is generated from a latent variable z through a generative process $p_\theta(x|z)$, where the latent variables follow a prior distribution $p(z)$, typically chosen as a standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The marginal likelihood of the data can be expressed as:

$$p_\theta(x) = \int p_\theta(x|z)p(z)dz \tag{3.30}$$

However, this integral is generally intractable, making direct maximum likelihood estimation infeasible. VAEs circumvent this challenge by introducing an approximate posterior $q_\phi(z|x)$, parameterized by a neural network (encoder), and optimizing a variational lower bound on the log-likelihood.

The **Evidence Lower BOund (ELBO)** is derived using Jensen's inequality:

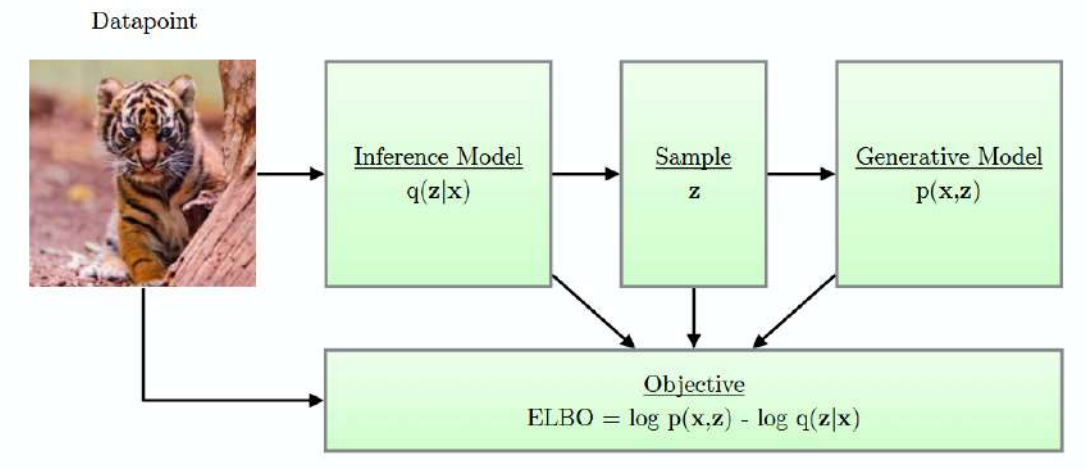


Figure 3.9. *Simple schematic of computational flow in a variational autoencoder. Source: [4]*

$$\log p_\theta(x) \geq \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)\|p(z)) \tag{3.31}$$
$$= \mathcal{L}(\theta, \phi; x) \tag{3.32}$$

The ELBO consists of two terms: a **reconstruction** term that encourages the decoder to accurately reconstruct the input from the latent representation, and a **regularization** term that constrains the approximate posterior to remain close to the prior distribution. This formulation enables simultaneous optimization of both the generative model $p_\theta(x|z)$ (decoder) and the inference model $q_\phi(z|x)$ (encoder).

The reparameterization trick

A critical innovation in VAEs is the reparameterization trick [56], which enables backpropagation through stochastic nodes. Instead of sampling directly from $q_\phi(z|x)$, the method expresses the random variable as a deterministic function of the parameters and an auxiliary noise variable. For a Gaussian posterior $q_\phi(z|x) = \mathcal{N}(\mu_\phi(x), \sigma^2_\phi(x))$, the reparameterization is:

$$z = \mu_\phi(x) + \sigma_\phi(x) \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{3.33}$$

where $\odot$ denotes the Hadamard (element-wise) product. This transformation allows gradients to flow through the sampling operation, enabling end-to-end optimization of the entire framework using standard backpropagation.

Training challenges

Posterior collapse occurs when the KL term dominates, causing the model to ignore the latent variable and reducing to a standard autoencoder. Multiple solutions, including KL/β-annealing [59], which devises a schedule for the KLD coefficient, and free bits [117],

which allow each latent variable a certain number of "free bits" without being penalized by the KLD term, have been proposed to address this issue.

Blurry reconstructions result in part due to the choice of reconstruction loss (typically ℓ_2) [60]. Advanced VAEs employ more sophisticated, perceptual loss functions or adversarial training (also analyzed in 3.4.2) to improve sample quality [61].

The **expressiveness gap** between the true posterior and the approximate posterior can limit model performance. To address this, extensive work has been done exploring the use of normalizing flows [62], a tool for constructing complex distributions by transforming a probability density through a sequence of invertible mappings, and inverse autoregressive flows [63], which improve performance compared to similar models with factorized Gaussian approximate posteriors.

Variants and extensions

Conditional VAEs (CVAEs) [118] extend the framework to conditional generation by incorporating additional information c (such as class labels) into both the encoder and decoder:

$$\mathcal{L}_{\text{CVAE}} = \mathbb{E}_{q_\phi(z|x,c)}[\log p_\theta(x|z,c)] - D_{KL}(q_\phi(z|x,c)\|p(z|c)) \tag{3.34}$$

β-VAE [58] introduces a hyperparameter β that weights the KL divergence term in the ELBO:

$$\mathcal{L}_{\beta\text{-VAE}} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \beta D_{KL}(q_\phi(z|x)\|p(z)) \tag{3.35}$$

Values of $\beta > 1$ encourage disentangled representations by imposing stronger regularization on the latent space, though often at the cost of reconstruction quality.

Vector Quantized VAEs (VQ-VAEs) [10] replace the continuous latent space with a discrete codebook, avoiding posterior collapse issues while enabling autoregressive modeling of discrete latent codes. The method introduces a quantization operation that maps encoder outputs to the nearest codebook vector.

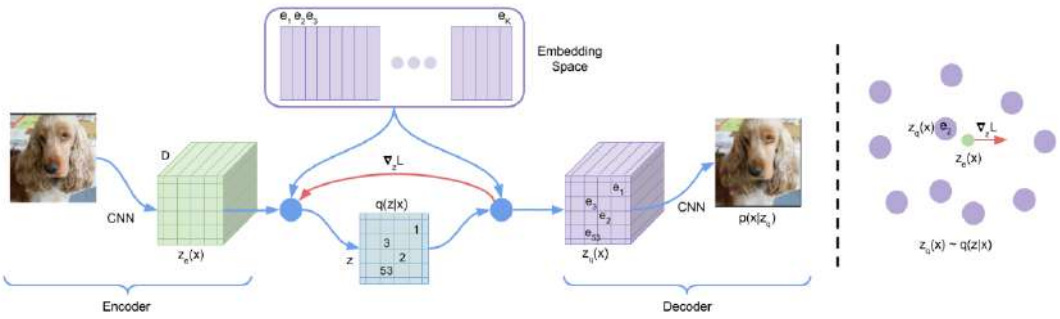


Figure 3.10. Left*: A figure describing the VQ-VAE.* Right*: Visualisation of the embedding space. The output of the encoder $z(x)$ is mapped to the nearest point e_2. The gradient $\nabla_z L$ (in red) will push the encoder to change its output, which could alter the configuration in the next forward pass. Source: [10]*

Nouveau VAE (NVAE) [119] is a deep hierarchical VAE which employs depth-wise separable convolutions and a residual parameterization of normal distributions. NVAE achieves training stability through spectral regularization.

3.4.2 Generative adversarial networks

Generative Adversarial Networks (GANs) [64] revolutionized generative modeling by introducing an adversarial training framework that pits two neural networks against each other in a minimax game. Unlike VAEs that rely on explicit likelihood computation, GANs learn to generate realistic samples through adversarial competition between a **generator** network that produces synthetic data and a **discriminator** network that distinguishes between real and generated samples.

Theoretical foundations

The GAN framework formulates generative modeling as a two-player minimax game. The generator $G_\theta : \mathbb{R}^d \to \mathbb{R}^n$ maps random noise vectors $z \sim p_z(z)$ to the data space, while the discriminator $D_\phi : \mathbb{R}^n \to [0, 1]$ estimates the probability that a given sample originates from the real data distribution rather than the generator. The training objective is:

$$\min_\theta \max_\phi V(D_\phi, G_\theta) = \mathbb{E}_{x \sim p_{\text{data}}}[\log D_\phi(x)] + \mathbb{E}_{z \sim p_z}[\log(1 - D_\phi(G_\theta(z)))] \tag{3.36}$$

At the optimal solution, the generator learns to approximate the data distribution p_{data} such that the discriminator cannot distinguish between real and generated samples, achieving $D^*(x) = 1/2$ for all x.

Training challenges

GAN training presents unique challenges due to its adversarial nature. The generator and discriminator must be carefully balanced—if one network becomes too strong relative to the other, training can become unstable or ineffective.

Mode collapse occurs when the generator learns to produce only a subset of the data distribution, failing to capture the full diversity of real samples [120]. This happens when the generator finds a few samples that consistently fool the discriminator, leading it to ignore other modes of the distribution.

Training instability appears as oscillatory behavior where the generator and discriminator alternate between different solutions without converging. The non-convex nature of the minimax objective and the discrete alternating optimization can lead to unstable gradients and training divergence [121].

Vanishing gradients can occur when the discriminator becomes too accurate early in training, providing uninformative gradients to the generator [122]. When $D(x) \approx 1$ for real data and $D(G(z)) \approx 0$ for generated data, the generator gradients vanish, preventing further learning.

Variants and extensions

Wasserstein GANs (WGANs) [11] address training instabilities by replacing the Jensen-Shannon divergence with the Wasserstein distance:

$$W_1(p_{\text{data}}, p_g) = \inf_{\gamma \sim \Pi(p_{\text{data}}, p_g)} \mathbb{E}_{(x,y)\sim\gamma}[\|x - y\|] \tag{3.37}$$

The WGAN objective becomes:

$$\min_{\theta} \max_{\phi:\|D_\phi\|_L \leq 1} \mathbb{E}_{x\sim p_{\text{data}}}[D_\phi(x)] - \mathbb{E}_{z\sim p_z}[D_\phi(G_\theta(z))] \tag{3.38}$$

where the discriminator (now called a critic) must satisfy a Lipschitz constraint. This formulation provides more stable training and meaningful loss curves that correlate with sample quality.

Gradient Penalty WGAN (WGAN-GP) [65] replaces weight clipping with a gradient penalty to enforce the Lipschitz constraint:

$$\mathcal{L}_{\text{GP}} = \mathbb{E}_{\hat{x}\sim p_{\hat{x}}}[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2] \tag{3.39}$$

where $\hat{x}$ represents points sampled along straight lines between real and generated samples.

Deep Convolutional GANs (DCGANs) [123] established architectural guidelines for stable GAN training, including the use of strided convolutions instead of pooling, batch normalization in both networks, and specific activation functions (ReLU in generator, LeakyReLU in discriminator).

Progressive GANs [124] enable generation of high-resolution images by gradually growing both generator and discriminator during training, starting from low resolution and progressively adding layers for higher resolution details.

StyleGAN [12] introduces style-based generation that provides unprecedented control over generated images through adaptive instance normalization:

$$\text{AdaIN}(x_i, y) = y_{s,i} \frac{x_i - \mu(x_i)}{\sigma(x_i)} + y_{b,i} \tag{3.40}$$

where $y = (y_{s,i}, y_{b,i})$ represents style vectors that control the mean and variance of feature maps at different scales.

3.4.3 Diffusion models

Diffusion models [66, 67], also known as score-based generative models, are a class of models grounded in the theory of stochastic differential equations (SDEs) that learn to reverse a gradual noise injection process [68]. The generative mechanism is conceptualized as the time-reversal of a forward diffusion process that progressively corrupts data until it is indistinguishable from pure Gaussian noise.

Algorithm 1 WGAN, our proposed algorithm. All experiments in the paper used the default values $\alpha = 0.00005$, $c = 0.01$, $m = 64$, $n_{\text{critic}} = 5$.

Require: : α, the learning rate. c, the clipping parameter. m, the batch size. n_{critic}, the number of iterations of the critic per generator iteration.
Require: : w_0, initial critic parameters. θ_0, initial generator's parameters.

while θ has not converged **do**
  for $t = 0, \ldots, n_{\text{critic}}$ **do**
    Sample $\{x^{(i)}\}_{i=1}^{m} \sim \mathbb{P}_r$ a batch from the real data.
    Sample $\{z^{(i)}\}_{i=1}^{m} \sim p(z)$ a batch of prior samples.
    $g_w \leftarrow \nabla_w \left[\frac{1}{m}\sum_{i=1}^{m} f_w(x^{(i)}) - \frac{1}{m}\sum_{i=1}^{m} f_w(g_\theta(z^{(i)}))\right]$
    $w \leftarrow w + \alpha \cdot \text{RMSProp}(w, g_w)$
    $w \leftarrow \text{clip}(w, -c, c)$
  end for
  Sample $\{z^{(i)}\}_{i=1}^{m} \sim p(z)$ a batch of prior samples.
  $g_\theta \leftarrow -\nabla_\theta \frac{1}{m}\sum_{i=1}^{m} f_w(g_\theta(z^{(i)}))$
  $\theta \leftarrow \theta - \alpha \cdot \text{RMSProp}(\theta, g_\theta)$
end while

Figure 3.11. *WGAN algorithm. Source: [11]*

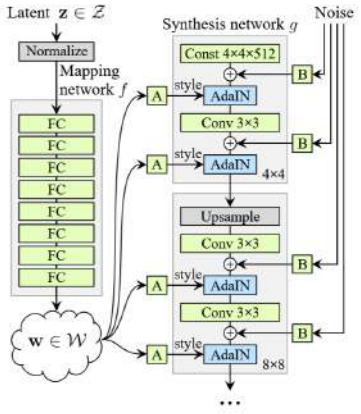


Figure 3.12. *Overview of StyleGan. Source: [12]*

Theoretical foundations

Forward diffusion process. The forward process is defined as a Markov chain that gradually adds Gaussian noise to the data x_0 over T discrete timesteps according to a fixed variance schedule $\{\beta_t\}_{t=1}^{T}$:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}). \tag{3.41}$$

A key property of this process is that the marginal distribution at any timestep t can be sampled in closed form without iterating through the chain. Letting $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, we have:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}). \tag{3.42}$$

This property is crucial for efficient training, as it allows for direct sampling of x_t for any t via the reparameterization: $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$ where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$.

Reverse diffusion process. The generative, or reverse, process aims to recover the original data by learning the conditional distributions $p_\theta(x_{t-1}|x_t)$. These are parameterized as Gaussians whose mean is predicted by a neural network $\mu_\theta(x_t, t)$:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2\mathbf{I}). \tag{3.43}$$

Starting from pure noise $x_T \sim \mathcal{N}(0, \mathbf{I})$, the model iteratively denoises the sample to produce a final estimate.

Training objective

While the model can be trained by maximizing the variational lower bound on the log-likelihood, Ho *et al.* [66] demonstrated that a simplified objective, which can be interpreted as a denoising score matching loss, yields superior results. The objective is to train a

network ϵ_θ to predict the noise component ϵ from the noisy sample x_t:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t\sim\mathcal{U}\{1,T\},x_0,\epsilon\sim\mathcal{N}(0,\mathbf{I})}\left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2\right]. \tag{3.44}$$

Score-based formulation

The continuous-time limit of diffusion models provides a unifying perspective through stochastic differential equations (SDEs) [69]. The forward process can be expressed as the solution to an SDE:

$$dx = f(x,t)dt + g(t)dw, \tag{3.45}$$

where $f(\cdot,t)$ is a drift and $g(t)$ a diffusion coefficient. The reverse of this process is also an SDE that depends on the score function of the data distribution, $\nabla_x \log p_t(x)$:

$$dx = [f(x,t) - g(t)^2\nabla_x \log p_t(x)]dt + g(t)d\bar{w}. \tag{3.46}$$

The neural network ϵ_θ is trained to approximate this score function (as introduced in [70]), as they are related by $\nabla_x \log p_t(x_t) \approx -\frac{\epsilon_\theta(x_t,t)}{\sqrt{1-\bar{\alpha}_t}}$.

This score-based framework gives rise to a corresponding deterministic process known as the Probability Flow ODE [69], which transports samples from the noise distribution to the data distribution by following the learned vector field. The ODE formulation is particularly useful for two reasons:

- **Fast sampling**: Numerical ODE solvers can approximate the trajectory with significantly fewer steps than stochastic samplers, generating high-quality samples in as few as 10-50 steps [71, 72].
- **Controllable generation**: The deterministic path provides a unique, invertible latent representation for each data point, which is valuable for tasks like image editing and interpolation [73].

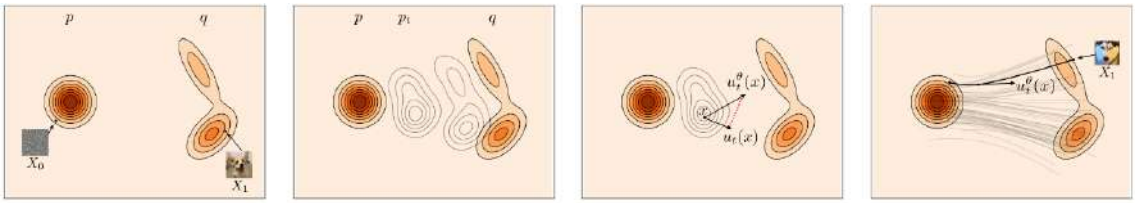
This score-based formulation reframes the generation process as learning a vector field that guides samples from a simple noise distribution back to the data manifold. This perspective of learning a vector field to transport densities is a powerful one, which has been recently generalized by Flow Matching [74], a central technique in this thesis.

3.4.4 Flow matching

Flow Matching (FM) [74, 5] constitutes a paradigm shift in continuous-time generative modeling by formulating the problem as learning vector fields that transport probability distributions. Unlike diffusion models that rely on stochastic processes, flow matching learns deterministic ordinary differential equations (ODEs) that map between distributions.

Theoretical foundations

FM assumes the existence of a time-dependent vector field $v : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$, implementing $v : (t,x) \mapsto v_t(x)$, that generates a flow $\phi : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$, implementing



(a) Data. (b) Path design. (c) Training. (d) Sampling.

Figure 3.13. *The Flow Matching blueprint. (a) The goal is to find a flow mapping samples X_0 from a known source or noise distribution p into samples X_1 from an unknown target or data distribution q. (b) To do so, design a time-continuous probability path $(p_t)_{0\leq t\leq 1}$ interpolating between $p := p_0$ and $q := p_1$. (c) During training, use regression to estimate the velocity field u_t known to generate p_t. (d) To draw a novel target sample $X_1 \sim q$, integrate the estimated velocity field $u_t^\theta(X_t)$ from $t = 0$ to $t = 1$, where $X_0 \sim p$ is a novel source sample. Source: [5]*

$\phi : (t, x) \mapsto \phi_t(x)$, which satisfies:

$$\frac{d\phi_t(x)}{dt} = v_t(\phi_t(x)), \quad \phi_0(x) = x. \tag{3.47}$$

The flow transports samples from a simple source distribution p_0 (typically Gaussian) at $t = 0$ to the target data distribution p_1 at $t = 1$. The evolution of probability densities follows the continuity equation:

$$\frac{\partial p_t}{\partial t} + \nabla \cdot (p_t v_t) = 0. \tag{3.48}$$

Conditional flow matching

Since learning the marginal vector field directly is intractable, FM constructs conditional flows $\phi_t(x|x_0, x_1)$ that interpolate between source points $x_0 \sim p_0$ and target points $x_1 \sim p_1$. As a result, the path and its corresponding vector field become simple and easy to define. For the commonly used **Optimal Transport (OT)** path, this is simply a straight line interpolation:

$$\phi_t(x_0, x_1) = (1 - t)x_0 + tx_1, \tag{3.49}$$

with the corresponding conditional vector field:

$$v_t(x_0, x_1) = \frac{d}{dt}\phi_t(x_0, x_1) = x_1 - x_0. \tag{3.50}$$

The marginal vector field is then:

$$v_t(x) = \mathbb{E}_{(x_0,x_1)\sim\pi}[v_t(x|x_0, x_1)|\phi_t(x_0, x_1) = x], \tag{3.51}$$

where π is a coupling between p_0 and p_1, *i.e.*, a joint probability distribution whose marginals along the respective axes are p_0 and p_1.

Training objective

The training objective simply minimizes the ℓ_2 distance between the learned vector field $v_t^\theta(x)$ and the target conditional vector field $v_t^*(x_t|x_0,x_1)$:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}[0,1],(x_0,x_1)\sim\pi}\left[\|v_t^\theta(x_t) - v_t^*(x_t|x_0,x_1)\|^2\right], \tag{3.52}$$

where $x_t = \phi_t(x_0,x_1)$. This is a simple regression objective that avoids the complexities of adversarial training or variational bounds. In the case of the OT path, substituting eq. 3.50 in eq. 3.52 yields:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}[0,1],(x_0,x_1)\sim\pi}\left[\|v_t^\theta(x_t) - (x_1 - x_0)\|^2\right]. \tag{3.53}$$

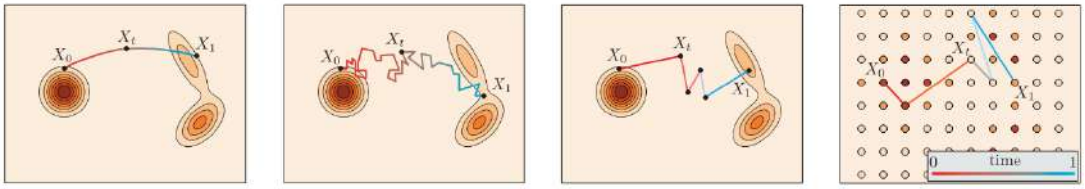
Stochastic interpolants

To ensure numerical stability and handle potential singularities, **stochastic interpolants (SIs)** [75, 76] are often employed:

$$\phi_t(x_0,x_1) = (1-t)x_0 + tx_1 + \sigma_t\epsilon, \tag{3.54}$$

where $\epsilon\sim\mathcal{N}(0,\mathbf{I})$ and σ_t is a noise schedule, typically $\sigma_t = \sigma\min(t,1-t)$ for some small $\sigma > 0$. The corresponding conditional vector field becomes:

$$v_t(x_t|x_0,x_1) = \frac{x_1 - (1-\sigma_t)x_0 - \sigma_t x_t}{\sigma_t} + \frac{d\sigma_t}{dt}\epsilon. \tag{3.55}$$



(a) Flow (b) Diffusion (c) Jump (d) CTMC

Figure 3.14. *Four time-continuous processes $(Xt)_{0\le t\le 1}$ taking source sample X_0 to a target sample X_1. These are **(a)** a flow in a continuous state space, **(b)** a diffusion in continuous state space, **(c)** a jump process in continuous state space (densities visualized with contours), and **(d)** a jump process in discrete state space (states as disks, probabilities visualized with colors). Source: [5]*

Advantages over traditional methods

1. **Training stability**: Unlike GANs, which require careful balancing of adversarial losses, FM optimizes a simple regression objective that is inherently stable and does not suffer from mode collapse.

2. **Exact likelihood**: FM provides exact likelihood computation without requiring expensive trace estimations, unlike traditional normalizing flows that rely on the

change of variables formula.

3. **High sample quality**: The continuous-time formulation enables smooth interpolation between distributions, often resulting in higher sample quality compared to discrete-step approaches.

4. **Computational efficiency**: The regression-based training objective as well the ODE formulation offer significant speedup, making FM suitable for large-scale applications.

Chapter 4

Methods

4.1 CrossFlowDG

Our previous navigation of the literature naturally leads us to the development of *CrossFlowDG*, a framework that addresses the modality gap using three main components: (1) a *Textual Domain Bank* (TDB), *i.e.*, a list of stylistic textual descriptions, (2) a *Four-way Contrastive Loss* (FCL) for intra- and inter-modal alignment, and (3) a *Cross-modal Flow Matching* (XFM) module that learns a deterministic mapping from domain-biased image to domain-invariant text representations. An overview of the framework is illustrated in Figure 4.1.

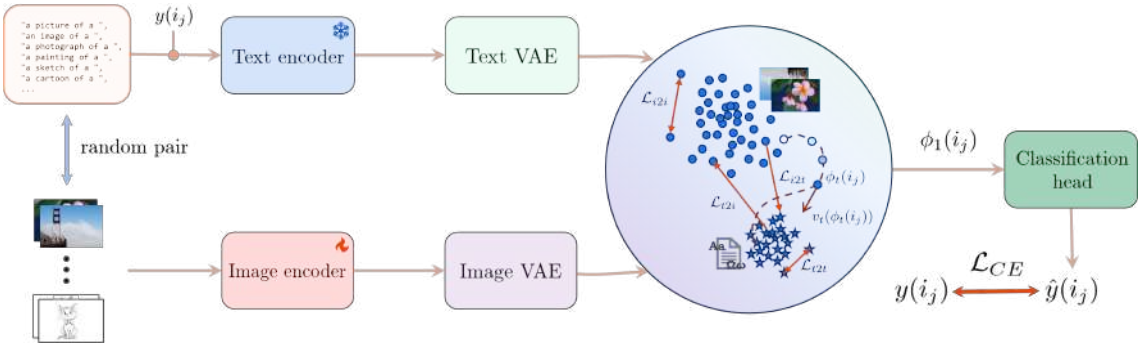


Figure 4.1. *Overview of the* CrossFlowDG *framework.*

4.1.1 Textual Domain Bank

In the realm of DG, a major challenge is obtaining supervision that encourages domain-invariant representations without requiring explicit domain labels or multiple source domains. To address this, we introduce a **Textual Domain Bank (TDB)**, *i.e.*, a collection of stylistic textual descriptions.

We construct the TDB as a set $\mathcal{D} = \{d_1, d_2, \ldots, d_K\}$ of K domain descriptors, following the template "a `{domain}` of a", where `domain` includes stylistic variations such as "photograph", "painting", "drawing" and similar terms capturing diverse visual styles. During training, each image x_i with class label y_i is paired with a text prompt constructed by randomly sampling a domain descriptor $d \sim \mathcal{U}(\mathcal{D})$ and concatenating it with the class name: $t_i = d + y_i$ ("+" as in string concatenation in Python). This produces descriptions

such as “a photograph of a dog” or “a sketch of a car”. Critically, the domain descriptor is sampled independently for each training iteration, meaning the same image may be paired with different domain descriptions across epochs.

This randomized pairing strategy serves a dual purpose. First, by exposing each image to multiple domain descriptors during training, the model learns that the same visual content can be described under various stylistic framings, encouraging the extraction of domain-invariant semantic features (specifically class information) that remain stable across these variations. Second, the distribution of text embeddings generated by the TDB for each class creates a semantically rich target space: rather than aligning images to a single fixed text embedding per class, images are encouraged to align with a manifold of textually-grounded semantic representations. This distributional view is important as it provides multiple semantic anchors per class, reducing the risk of overfitting to domain-specific visual patterns while preserving the core semantic content that generalizes across domains.

4.1.2 Four-way Contrastive Loss

To enforce domain invariance in image representations through alignment with the TDB, we employ a **Four-way Contrastive Loss (FCL)** consisting of multiple complementary loss terms: (1) image-to-image ($\mathcal{L}_{i2i}$), to compress the intra-class variance across domains, creating compact same-class image clusters, (2) text-to-text ($\mathcal{L}_{t2t}$), with the same goal as (1) but for the text representations, and (3) cross-modal ($\mathcal{L}_{i2t}$ and $\mathcal{L}_{t2i}$), to bridge the modality gap by enforcing alignment between image and text embeddings of the same class. $\mathcal{L}_{i2t}$ treats each image as positive with its corresponding text (from the TDB), pulling image representations toward the domain-invariant text manifold, while $\mathcal{L}_{t2i}$ enforces the symmetric constraint.

In order to apply the contrastive objective, we project our extracted features to a Euclidean shared latent space using two separate VAEs, one for each modality. Let f^i denote the pretrained image encoder, g^i the image VAE, f^t the pretrained text encoder, and g^t the text VAE. $z_j^i = g^i(f^i(x_j))$ denotes the latent image representation and $z_j^t = g^t(f^t(t_j))$ the corresponding text latent. For the computation of the loss, features are normalized to the unit hypersphere by h, such that $\|z_j\|_2 = 1$. We define cosine similarity as:

$$s(z_a, z_b) = \frac{h(z_a)^\top h(z_b)}{\tau}, \tag{4.1}$$

where $\tau > 0$ is a temperature parameter. Each contrastive loss term between modalities k and l is then defined as:

$$\mathcal{L}_{k2l} = -\frac{1}{N}\sum_{j=1}^{N} \log \frac{\exp(s(z_j^k, z_j^l))}{\sum_{m=1}^{N} \exp(s(z_j^k, z_m^l))}. \tag{4.2}$$

The total contrastive loss is:

$$\mathcal{L}_{\text{FC}} = \mathcal{L}_{i2t} + \mathcal{L}_{t2i} + \frac{1}{2}(\mathcal{L}_{t2t} + \mathcal{L}_{i2i}). \tag{4.3}$$

This objective encourages class-wise cross-modal alignment while preserving intra-modal cohesion.

4.1.3 Cross-modal Flow Matching

As stated previously, standard cosine similarity contrastive learning does not entirely close the modality gap. We therefore employ noise-free, **Cross-modal Flow Matching (XFM)** to explicitly learn a continuous transformation that maps image latents to text latents while preserving the intrinsic geometry of the latent space. Unlike conventional generative flows that map noise to data, we adopt a deterministic cross-modal flow that connects image and text manifolds in latent space, similar to [77].

Given paired image-text latents $(z_{\text{img}}, z_{\text{txt}})$, we define an interpolated sample at time $t \in [0,1]$:

$$z_t = (1-t)z_{\text{img}} + t z_{\text{txt}}. \tag{4.4}$$

The ground-truth velocity field is:

$$v^*(z_t, t) = z_{\text{txt}} - z_{\text{img}}. \tag{4.5}$$

A neural flow model that predicts the velocity $v^\theta(z_t, t)$ is trained via:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}[0,1],(z_{\text{img}},z_{\text{txt}})\sim\pi_{\text{rand}}}\left[\|v_t^\theta(z_t) - v_t^*(z_t|z_{\text{img}}, z_{\text{txt}})\|^2\right], \tag{4.6}$$

where π_{rand} is the random mini-batch image-text coupling described in 4.1.1.

At inference, the learned flow defines a deterministic mapping from image to text latent:

$$z_1 = z_{\text{img}} + \int_0^1 v^\theta(z_t, t)\, dt, \tag{4.7}$$

which is approximated using a small number of discrete integration steps:

$$z_1 \approx z_{\text{img}} + \sum_{i=0}^{T-1} v^\theta(z_{\frac{i}{T}}, \frac{i}{T})\frac{1}{T}. \tag{4.8}$$

This allows the model to align distributions while preserving the topology of the shared latent space, facilitating domain-invariant generalization.

4.2 OT-CrossFlowDG

4.2.1 Motivation

Despite the benefits of *CrossFlowDG*, our goal of enhancing efficiency during inference, and thus facilitating deployability on devices with computational limitations, points to the direction of the flow model. It should be noted that *CrossFlowDG* uses 12 discrete timesteps to approximate the ODE integration, which means that the flow model has to be queried 12 times before delivering the domain-invariant embedding at $t = 1$. In an attempt to reduce the number of ODE integration steps, we devise Algorithm 4.2 to

rectify our naive, random image-text coupling and bootstrap our conditional flow matching formulation with optimal transport-informed pairs, similar to [78].

4.2.2 Class-wise Sinkhorn-Knopp coupling algorithm

Algorithm 4.2 initially computes the entropic regularized optimal transport (OT) distance between a set of image embeddings $\mathbf{E}_I \in \mathbb{R}^{B\times D}$ and text embeddings $\mathbf{E}_T \in \mathbb{R}^{K\times D}$ while enforcing a strict class-based alignment constraint. Initially, a dense cost matrix $\mathbf{C}$ is computed using the **squared Euclidean distance** between all image-text pairs of a mini-batch (which means that $B = K$ in our case). Then, a large penalty cost L is applied to all cross-class pairs, effectively masking the problem to permit coupling only between image-text instances that share the same ground-truth label. The resulting constrained cost matrix $\mathbf{C}_{\text{mask}}$ is fed to the **Sinkhorn-Knopp** [79] algorithm (described in Algorithm 4.1 for ease of reference) in the log-domain for numerical stability, which yields the approximate optimal transport plan $\mathbf{P}$. The final scalar loss $\mathcal{L}_{\text{CW}}$ is defined as the expected cost under the plan $\mathbf{P}$ summed only over the valid, same-class couplings. A key secondary output of this process is the set of **barycentric targets** $\mathbf{B} \in \mathbb{R}^{K\times D}$, which are calculated as the weighted mean of the text embeddings according to the learned transport plan $\mathbf{P}$. The use of barycentric targets serves two main purposes: (1) to simulate sampling diverse text embeddings from the empirical (mini-batch) distribution, and (2) to reduce the influence of potential outliers on the flow matching process.

Algorithm 4.1: *Sinkhorn-Knopp [79]*

Require: cost matrix $\mathbf{C} \in \mathbb{R}^{n\times m}$, entropic regularization $\epsilon > 0$, T iterations.
Ensure: approximate optimal transport plan $\mathbf{P} \in \mathbb{R}^{n\times m}$.

$\boldsymbol{\mu} \leftarrow \mathbf{1}_n/n$ ▷ uniform marginal for source
$\boldsymbol{\nu} \leftarrow \mathbf{1}_m/m$ ▷ uniform marginal for target
$\mathbf{u} \leftarrow \mathbf{0}_n, \mathbf{v} \leftarrow \mathbf{0}_m$ ▷ initialize potentials in log-domain
for $t \leftarrow 1$ to T **do**
  $\mathbf{u}_i \leftarrow -\epsilon \log\left(\sum_{j=1}^{m} \exp\left(\frac{-C_{ij}+v_j}{\epsilon}\right)\right) + \epsilon \log(\mu_i)$ ▷ K-M scaling: $\mathbf{u} \leftarrow \mathbf{u} - \log(\mathbf{Kv})$
  $\mathbf{v}_j \leftarrow -\epsilon \log\left(\sum_{i=1}^{n} \exp\left(\frac{-C_{ij}+u_i}{\epsilon}\right)\right) + \epsilon \log(\nu_j)$ ▷ K-M scaling: $\mathbf{v} \leftarrow \mathbf{v} - \log(\mathbf{K}^T\mathbf{u})$
end for
$\mathbf{P}_{ij} \leftarrow \exp\left(\frac{-C_{ij}+u_i+v_j}{\epsilon}\right)$ ▷ final transport plan
return P

Algorithm 4.2: *Class-wise Sinkhorn-Knopp coupling*

Require: image embeddings $\mathbf{E}_I \in \mathbb{R}^{B\times D}$, text embeddings $\mathbf{E}_T \in \mathbb{R}^{K\times D}$,
Require: image labels $\mathbf{L}_I \in \mathbb{R}^{B}$, text labels $\mathbf{L}_T \in \mathbb{R}^{K}$,
Require: entropic regularization $\epsilon > 0$, large penalty cost $L > 0$, T Sinkhorn-Knopp iterations.
Ensure: total loss $\mathcal{L}_{\text{CW}}$, barycentric targets $\mathbf{B}$.

▷ 1. Compute cost matrix $\mathbf{C}$
$\mathbf{C} \leftarrow \text{PairwiseSquaredDistance}(\mathbf{E}_I, \mathbf{E}_T)$
▷ 2. Mask $\mathbf{C}$ to enforce class constraint
$\mathbf{M}_{ij} \leftarrow (\mathbf{L}_{I,i} == \mathbf{L}_{T,j})$ ▷ binary mask for same-class pairs
$\mathbf{C}_{\text{mask}} \leftarrow \mathbf{C} \odot \mathbf{M} + (\mathbf{I} - \mathbf{M}) \cdot L$ ▷ apply large penalty L to cross-class pairs
▷ 3. Compute transport plan
$\mathbf{P} \leftarrow \text{Sinkhorn-Knopp}(\mathbf{C}_{\text{mask}}, \epsilon, T)$ ▷ run Sinkhorn-Knopp iterations (Algorithm 4.1)
▷ 4. Compute loss (expected cost)
$\mathcal{L}_{\text{CW}} \leftarrow \frac{\sum_{i,j} \mathbf{P}_{ij} \cdot \mathbf{C}_{ij} \cdot \mathbf{M}_{ij}}{\sum_{i,j} \mathbf{M}_{ij}}$ ▷ loss is calculated only over same-class pairs
▷ 5. Compute barycentric targets
$\mathbf{P}_{\text{masked}} \leftarrow \mathbf{P} \odot \mathbf{M}$
$\mathbf{S}_i \leftarrow \sum_j (\mathbf{P}_{\text{masked}})_{ij}$ ▷ row sums of the masked plan
$\mathbf{B}_i \leftarrow \frac{\sum_j (\mathbf{P}_{\text{masked}})_{ij} \cdot (\mathbf{E}_T)_j}{\mathbf{S}_i}$ ▷ barycentric mean of text embeddings

return $\mathcal{L}_{\text{CW}}, \mathbf{B}$

Chapter 5

Experiments

5.1 Benchmarks

We evaluate classification accuracy on four usual DG image datasets, *i.e.*, **TerraIncognita** [80], **PACS** [81], **VLCS** [82], and **OfficeHome** [83], adopting the *leave-one-domain-out* evaluation protocol: if the dataset contains N domains $\mathcal{D}_i, i \in \{1, 2, \ldots, N\}$, we train on $N-1$ domains, evaluate on the remaining one, repeat for each test domain and compute the average:

$$\text{LODO}_{\mathcal{D}} = \frac{1}{N} \sum_{i=1}^{N} R_{\setminus i}, \tag{5.1}$$

where $R_{\setminus i}$ denotes the model's accuracy when trained on $\mathcal{D} \setminus \{\mathcal{D}_i\}$.

For fair comparison with previous methods, we train for 50 epochs with 200 gradient updates per epoch (10,000 iterations in total). Reported accuracy for *CrossFlowDG* is explicitly computed as an average across 3 distinct random seeds along with the standard deviation, contrary to some prior work. Due to resource constraints, for *OT-CrossFlowDG* we perform 3 runs, each with a different number of ODE integration steps to outline any potential difference in performance.

Specifically, the **TerraIncognita** dataset contains 24330 images harvested from camera traps deployed in 4 different locations, which depict animals of 10 different species. **PACS** consists of 9991 images belonging to 7 common object categories and mainly focuses on large shifts in visual style. **VLCS** includes 10729 samples collected from four popular and distinct source image datasets and provides a benchmark on 5 common classes ("*bird*", "*car*", "*chair*", "*dog*", and "*person*"). **OfficeHome** comprises 15588 images of 65 everyday objects typically found in office and home environments (*e.g.*, "*Alarm-Clock*", "*Bed*", "*Chair*", "*Mug*", etc.). Table 5.1 offers a quantitative summary of the above four datasets.

5.2 Baselines

To perform a comprehensive evaluation, we compare our proposed model against several SOTA architectures used as backbones for image classification in the context of DG. We specifically select models that are similar in parameter count and use the widely adopted ImageNet-1K [86] pretraining for fair comparison.

Table 5.1. *A quantitative view of the four benchmarks used for evaluation.*

Dataset	Domain	# images	# images	# classes
TerraIncognita	L100	4741	24330	10
	L38	9736		
	L43	3970		
	L46	5883		
PACS	Art painting	2048	9991	7
	Cartoon	2344		
	Photo	1670		
	Sketch	3929		
VLCS	Caltech101	1415	10729	5
	LabelMe	2656		
	SUN09	3282		
	VOC2007	3376		
OfficeHome	Art	2427	15588	65
	Clipart	4365		
	Product	4439		
	Real	4357		

Our selected baseline backbones include:

- **ResNet-50** [84]: A classic Convolutional Neural Network (CNN) that introduced residual connections, serving as a robust and widely accepted standard for many vision tasks.
- **DeiT-S** [85]: A Vision Transformer (ViT) variant optimized for efficiency. This backbone represents is particularly relevant for demonstrating performance relative to models utilizing the attention mechanism.
- **VMamba-T** [9]: A lightweight variant of the VMamba architecture. It utilizes the State Space Model (SSM) for efficient sequence modeling in vision, offering a strong baseline against both CNNs and Transformers.

The similar parameter count ensures that any observed performance differences are likely not attributed to model size, providing a clear and reliable set of baselines.

5.3 Implementation details

5.3.1 Training objective

The overall training objective for *CrossFlowDG* combines the VAE, contrastive and flow components as defined in Section 4.1:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{VAE}}^{\text{img}}\mathcal{L}_{\text{VAE}}^{\text{img}} + \lambda_{\text{VAE}}^{\text{txt}}\mathcal{L}_{\text{VAE}}^{\text{txt}} + \lambda_{\text{FC}}\mathcal{L}_{\text{FC}} + \lambda_{\text{FM}}\mathcal{L}_{\text{FM}}, \tag{5.2}$$

where $\mathcal{L}^k_{\text{VAE}} = \mathcal{L}^k_{\text{recon}} + \mathcal{L}^k_{\text{KL}}$ is the VAE loss that corresponds to modality k and includes a reconstruction term (*e.g.*, mean squared error loss) and a regularization term (*e.g.*, Kullback-Leibler divergence between the distribution of the latent representation and a standard Gaussian), whereas the λ_i coefficients control the contributions of the respective terms. For *OT-CrossFlowDG*, $\mathcal{L}_{\text{FC}}$ is simply replaced by $\mathcal{L}_{\text{CW}}$ as described in Section 4.2.

All trainable components of our methods are trained jointly. Also, all trainable components are trained from scratch, except for the pretrained image encoder.

5.3.2 Network architecture

Note: "M" below denotes "million".

- **Image encoder:** VMamba-T [9], 29M trainable parameters.
- **Text encoder:** CLIP's text encoder [49], no trainable parameters.
- **VAE:** Multi-Layer Perceptron (MLP) encoder and decoder, 2 VAEs used, with 4M and 1M trainable parameters.
- **Flow model:** ResNet [84], 1M trainable parameters.
- **Classification head:** MLP, 0.02M trainable parameters.

Our proposed method comprises 36M trainable parameters in total.

5.3.3 Latent space

- **Latent dimension:** 256
- **ODE integration steps:** *CrossFlowDG*: 12, *OT-CrossFlowDG*: 1, 6, 12

5.3.4 Training details

- **Batch size:** 16 (per source domain)
- **Iterations:** 10,000 total iterations (50 epochs × 200 gradient updates / epoch)
- **Hardware:** 1 × A10G GPU

5.4 Results

5.4.1 CrossFlowDG

Results on TerraIncognita. Table 5.2 clearly illustrates the substantial performance gain of CrossFlowDG on the TerraIncognita benchmark compared to the previous SOTA method. We observe an improvement ranging from 2.2% to 3.5% across three out of the four target domains, resulting in a notable +1.9% average accuracy gain.

Results on PACS. *CrossFlowDG*'s performance on PACS is presented in Table 5.3. *CrossFlowDG* achieves the highest accuracy (compared to previous methods) in the 'art painting' domain, the second-highest in the 'cartoon' and 'photo' domains and the second-highest on average (by -0.5%).

Results on VLCS. *CrossFlowDG* achieves SOTA in one domain of VLCS, while its average performance remains 1.2% below SOTA, as is visible in Table 5.4.

Results on OfficeHome. As shown in Table 5.5, CrossFlowDG performs 2.9% below the previous SOTA method on average (with a minimum difference of 2.5% and a maximum difference of 3.3% across domains).

5.4.2 OT-CrossFlowDG

For brevity, we report the best performing (on average) ResNet-50 method, the best performing (on average) DeiT-S method, the two VMamba-based methods and our own results for 1, 6 and 12 ODE integration steps.

In general, *OT-CrossFlowDG* performs similarly to *CrossFlowDG* on average, except for TerraIncognita (Table 5.6), where we observe a sizeable degradation (54.4% for T=1 versus 58.0% for *CrossFlowDG*). As stated in Section 4.2, a primary objective of *OT-CrossFlowDG* is to increase efficiency, and an observation of performance relative to the number of ODE integration steps suggests the existence of a trend: *OT-CrossFlowDG* generally exhibits its peak performance at a low number of integration steps. However, an exception to this pattern is observed in the OfficeHome benchmark, where performance seems to improve as the number of steps increases.

Table 5.2. *Performance comparison on TerraIncognita.* ***Bold*** *and* <u>underline</u> *indicate best and second-best performance, respectively.*

Method	Venue	Param's	Target domain				Avg. (↑)
			L100	L38	L43	L46	
ResNet-50 based:							
GroupDRO	ICLR '19	23M	41.2	38.6	56.7	36.4	43.2
VReX	ICML '21	23M	48.2	41.7	56.8	38.7	46.4
RSC	ECCV '20	23M	50.2	39.2	56.3	40.8	46.6
MTL	JMLR '21	23M	49.3	39.6	55.6	37.8	45.6
Mixstyle	ICLR '21	23M	54.3	34.1	55.9	31.7	44.0
SagNet	CVPR '21	23M	53.0	43.0	57.9	40.4	48.6
ARM	NeurIPS '21	23M	49.3	38.3	55.8	38.7	45.5
SWAD	NeurIPS '21	23M	55.4	44.9	59.7	39.9	50.0
PCL	CVPR '22	23M	58.7	46.3	60.0	43.6	52.1
SAGM	CVPR '23	23M	54.8	41.4	57.7	41.3	48.8
iDAG	ICCV '23	23M	58.7	35.1	57.5	33.0	46.1
GMDG	CVPR '24	23M	59.8	45.3	57.1	38.2	50.1
DeiT-S based:							
SDViT	ACCV '22	22M	55.9	31.7	52.2	37.4	44.3
GMoE	ICLR '23	34M	59.2	34.0	50.7	38.5	45.6
VMamba-T based:							
DGMamba	MM '24	31M	62.0	47.7	61.7	46.9	54.5
DGFamba	AAAI '25	31M	<u>63.9</u>	<u>49.8</u>	**63.1**	<u>47.5</u>	<u>56.1</u>
CrossFlowDG (1)	2025	36M	69.3	50.4	64.4	50.9	58.7
CrossFlowDG (2)	2025	36M	62.5	49.2	63.6	49.9	56.3
CrossFlowDG (3)	2025	36M	68.0	56.3	60.3	50.4	58.8
CrossFlowDG (avg.)	2025	36M	**66.6** (±3.6)	**52.0** (±3.8)	<u>62.8</u> (±2.2)	**50.4** (±0.5)	**58.0** (±1.4)

Table 5.3. *Performance comparison on PACS.* ***Bold*** *and* <u>underline</u> *indicate best and second-best performance, respectively.*

Method	**Target domain**				**Avg.** (↑)
	A	**C**	**P**	**S**	
ResNet-50 based:					
GroupDRO	83.5	79.1	96.7	78.3	84.4
VReX	86.0	79.1	96.9	77.7	84.9
RSC	85.4	79.7	97.6	78.2	85.2
MTL	87.5	77.1	96.4	77.3	84.6
Mixstyle	86.8	79.0	96.6	78.5	85.2
SagNet	87.4	80.7	97.1	80.0	86.3
ARM	86.8	76.8	97.4	79.3	85.1
SWAD	89.3	83.4	97.3	82.5	88.1
PCL	90.2	83.9	98.1	82.6	88.7
SAGM	87.4	80.2	98.0	80.8	86.6
iDAG	90.8	83.7	98.0	82.7	88.8
GMDG	84.7	81.7	97.5	80.5	85.6
DeiT-S based:					
SDViT	87.6	82.4	98.0	77.2	86.3
GMoE	89.4	83.9	99.1	74.5	86.7
VMamba-T based:					
DGMamba	91.3	87.0	99.0	<u>87.3</u>	91.2
DGFamba	<u>92.6</u>	**89.4**	**99.7**	**88.8**	**92.6**
CrossFlowDG (1)	93.2	87.6	99.5	86.8	91.8
CrossFlowDG (2)	93.3	89.9	99.4	86.3	92.2
CrossFlowDG (3)	93.3	89.4	99.4	87.3	92.4
CrossFlowDG (avg.)	**93.3** (±0.1)	<u>89.0</u> (±1.2)	<u>99.4</u> (±0.1)	86.8 (±0.5)	<u>92.1</u> (±0.3)

Table 5.4. *Performance comparison on VLCS.* ***Bold*** *and* <u>underline</u> *indicate best and second-best performance, respectively.*

Method	Target domain				Avg. (↑)
	C	L	S	P	
ResNet-50 based:					
GroupDRO	97.3	63.4	69.5	76.7	76.7
VREx	98.4	64.4	74.1	76.2	78.3
RSC	97.9	62.5	72.3	75.6	77.1
MTL	97.8	64.3	71.5	75.3	77.2
Mixstyle	98.6	64.5	72.6	75.7	77.9
SagNet	97.9	64.5	71.4	77.5	77.8
ARM	98.7	63.6	71.3	76.7	77.6
SWAD	98.8	63.3	75.3	79.2	79.1
PCL	<u>99.0</u>	63.6	73.8	75.6	78.0
SAGM	<u>99.0</u>	65.2	75.1	80.7	80.0
iDAG	98.1	62.7	69.9	77.1	76.9
GMDG	98.3	65.9	73.4	79.3	79.2
DeiT-S based:					
SDViT	96.8	64.2	76.2	78.5	78.9
GMoE	96.9	63.2	72.3	79.5	78.0
VMamba-T based:					
DGMamba	98.9	64.3	79.2	<u>80.8</u>	80.8
DGFamba	**99.5**	**66.2**	<u>80.9</u>	**82.0**	**82.2**
CrossFlowDG (1)	98.0	66.3	80.7	81.2	81.5
CrossFlowDG (2)	96.5	65.9	81.3	77.7	80.3
CrossFlowDG (3)	97.6	66.1	81.8	79.7	81.3
CrossFlowDG (avg.)	97.3 (±0.6)	<u>66.1</u> (±0.2)	**81.3** (±0.5)	79.5 (±1.5)	<u>81.0</u> (±0.6)

Table 5.5. *Performance comparison on OfficeHome.* ***Bold*** *and* underline *indicate best and second-best performance, respectively.*

Method	Target domain				Avg. (↑)
	A	**C**	**P**	**R**	
ResNet-50 based:					
GroupDRO	60.4	52.7	75.0	76.0	66.0
VREx	60.7	53.0	75.3	76.6	66.4
RSC	60.7	51.4	74.8	75.1	65.5
MTL	61.5	52.4	74.9	76.8	66.4
Mixstyle	51.1	53.2	68.2	69.2	60.4
SagNet	63.4	54.8	75.8	78.3	68.1
ARM	58.9	51.0	74.1	75.2	64.8
SWAD	66.1	57.7	78.4	80.2	70.6
PCL	67.3	59.9	78.7	80.7	71.6
SAGM	65.4	57.0	78.0	80.0	70.1
iDAG	68.2	57.9	79.7	81.4	71.8
GMDG	68.9	56.2	79.9	82.0	70.7
DeiT-S based:					
SDViT	68.3	56.3	79.5	81.8	71.5
GMoE	69.3	58.0	79.8	82.6	72.4
VMamba-T based:					
DGMamba	<u>76.2</u>	<u>61.8</u>	<u>83.9</u>	<u>86.1</u>	<u>77.0</u>
DGFamba	**77.4**	**63.7**	**85.6**	**87.3**	**78.5**
CrossFlowDG (1)	74.8	60.7	82.3	83.8	75.4
CrossFlowDG (2)	74.4	61.2	82.7	83.6	75.5
CrossFlowDG (3)	75.6	60.9	82.8	84.6	76.0
CrossFlowDG (avg.)	74.9 (± 0.6)	60.9 (± 0.3)	82.6 (± 0.3)	84.0 (± 0.5)	75.6 (± 0.3)

Table 5.6. *Performance comparison on TerraIncognita.* ***Bold*** *and* <u>underline</u> *indicate best and second-best performance, respectively.*

Method	Target domain				Avg. (↑)
	L100	**L38**	**L43**	**L46**	
PCL	58.7	46.3	60.0	43.6	52.1
GMoE	59.2	34.0	50.7	38.5	45.6
DGMamba	62.0	<u>47.7</u>	<u>61.7</u>	46.9	<u>54.5</u>
DGFamba	<u>63.9</u>	**49.8**	**63.1**	47.5	**56.1**
OT-CrossFlowDG ($T = 1$)	63.2	45.9	59.3	**49.3**	54.4
OT-CrossFlowDG ($T = 6$)	**65.9**	47.0	57.0	46.6	54.1
OT-CrossFlowDG ($T = 12$)	62.0	45.6	57.0	<u>48.7</u>	53.3

Table 5.7. *Performance comparison on PACS.* ***Bold*** *and* <u>underline</u> *indicate best and second-best performance, respectively.*

Method	Target domain				Avg. (↑)
	A	**C**	**P**	**S**	
iDAG	90.8	83.7	98.0	82.7	88.8
GMoE	89.4	83.9	99.1	74.5	86.7
DGMamba	91.3	87.0	99.0	<u>87.3</u>	91.2
DGFamba	92.6	**89.4**	**99.7**	**88.8**	**92.6**
OT-CrossFlowDG ($T = 1$)	93.6	88.0	<u>99.5</u>	86.0	<u>91.8</u>
OT-CrossFlowDG ($T = 6$)	<u>93.7</u>	<u>89.2</u>	99.3	85.0	<u>91.8</u>
OT-CrossFlowDG ($T = 12$)	**94.1**	88.0	99.2	85.2	91.6

Table 5.8. *Performance comparison on VLCS.* ***Bold*** *and* <u>underline</u> *indicate best and second-best performance, respectively.*

Method	Target domain				Avg. (↑)
	C	**L**	**S**	**P**	
SAGM	<u>99.0</u>	65.2	75.1	80.7	80.0
SDViT	96.8	64.2	76.2	78.5	78.9
DGMamba	98.9	64.3	79.2	<u>80.8</u>	80.8
DGFamba	**99.5**	66.2	80.9	**82.0**	**82.2**
OT-CrossFlowDG ($T = 1$)	97.8	**66.7**	**81.6**	79.4	<u>81.4</u>
OT-CrossFlowDG ($T = 6$)	97.4	<u>66.5</u>	<u>81.0</u>	80.3	81.3
OT-CrossFlowDG ($T = 12$)	97.2	65.6	<u>81.0</u>	78.6	80.6

Table 5.9. *Performance comparison on OfficeHome.* ***Bold*** *and* underline *indicate best and second-best performance, respectively.*

Method	Target domain				Avg. (↑)
	A	**C**	**P**	**R**	
iDAG	68.2	57.9	79.7	81.4	71.8
GMoE	69.3	58.0	79.8	82.6	72.4
DGMamba	76.2	61.8	83.9	86.1	77.0
DGFamba	**77.4**	**63.7**	**85.6**	**87.3**	**78.5**
OT-CrossFlowDG ($T = 1$)	74.4	60.9	82.6	84.2	75.5
OT-CrossFlowDG ($T = 6$)	75.4	61.7	82.5	84.5	76.0
OT-CrossFlowDG ($T = 12$)	74.1	61.0	82.9	84.9	75.7

Chapter 6

Discussion

6.1 Conclusions

In this thesis, we introduce two novel methods to tackle the critical problem of image classification under domain generalization (DG): *CrossFlowDG* and *OT-CrossFlowDG*, the latter being an optimal transport-informed variant of the former.

The first proposed framework, *CrossFlowDG*, initiates image-text alignment with a preliminary cosine-based contrastive loss in a Euclidean joint latent space, followed by our flow matching mechanism, which addresses the modality gap by learning a continuous transformation between the partially aligned image and text embeddings of the same semantic class.

From a performance-centric perspective, *CrossFlowDG* achieves state-of-the-art results across several domains of our evaluation datasets, most notably on TerraIncognita, a benchmark we consider the most challenging both quantitatively and qualitatively. Its high difficulty is evidenced by the consistently lowest accuracy scores achieved by previous methods, which barely exceed 54% as shown in Table 5.2. Qualitatively, the images, captured via camera traps in uncontrolled natural environments, feature severe domain shifts characterized by motion blur and low-light conditions, which greatly hinder accurate classification even by human annotators.

Conversely, near-perfect results on certain domains highlight the practical limits of improvement. Across all listed methods, the 'photo' domain of PACS and the 'Caltech101' domain of VLCS exhibit highly saturated performance, with accuracies in most recent methods exceeding 99.0%. This high saturation likely reflects that modern backbones have been extensively pre-trained on samples similar (or even identical) to those contained in PACS and VLCS. Consequently, marginal differences in these domains are potentially less indicative of a model's true generalization capacity.

Furthermore, *CrossFlowDG* demonstrates suboptimal performance across the four domains of OfficeHome. We hypothesize that the primary reason for this degradation is the dataset's large label space. Specifically, OfficeHome contains 65 classes, a factor which significantly hinders the formation of well-separated and distinct class clusters in the latent space. This lack of clearly separated clusters subsequently reduces the effectiveness of our flow matching mechanism, which fundamentally relies on learning continuous paths between the image distribution and the semantically consistent text distribution.

The second proposed framework, *OT-CrossFlowDG*, emphasizes efficient inference, with the core differences from *CrossFlowDG* being twofold: (1) the cross-modal loss is computed using the entropic regularized 2-Wasserstein distance, and (2) the Sinkhorn-Knopp algorithm is applied to derive barycentric targets from an approximate optimal transport plan. This optimal transport-informed coupling yields a highly informative supervision signal that significantly reduces the number of integration steps required by our flow matching formulation during inference.

Regarding the experimental results of *OT-CrossFlowDG*, we observe that OfficeHome represents a notable exception to the trend of achieving peak performance with fewer integration steps. When the latent space must accommodate OfficeHome's 65 classes, the resulting class clusters become tightly packed and potentially overlapping. To accurately navigate this complex space, a finer discretization of the flow trajectory, *i.e.*, a higher number of integration steps, is likely required to maintain classification accuracy.

6.2 Limitations and future work

Despite the advantages of our proposed methods, we acknowledge some key limitations that may serve as a compass for future research:

Combining training stability with inference efficiency. *CrossFlowDG* enjoys stable training due to the use of cosine similarity, whereas *OT-CrossFlowDG* can be computationally advantageous during inference but tends to suffer from vanishing or unstable gradients, perhaps due to the high flexibility introduced by Euclidean distances in the OT cost matrix. Future work could combine the benefits of both approaches by employing cosine-based distances within the OT framework.

Generality of proposed frameworks. We evaluated our methods using only VMamba-T and CLIP's text encoder as backbones. Further experimentation with diverse combinations of image and text encoders is needed to provide stronger empirical evidence for the generality of our frameworks. Notably, a key feature of our approach is that it does not rely on perfectly aligned modalities from the backbone pretraining stage; instead, it explicitly enforces alignment during training. Consequently, *CrossFlowDG* and *OT-CrossFlowDG* could be extended to scenarios involving an arbitrary number of modalities, enabling more flexible and comprehensive cross-modal alignment.

Evaluation on saturated benchmarks. As discussed, certain domains exhibit performance saturation over 99%. These saturated results may reflect data leakage from pretraining rather than genuine domain generalization capabilities. Future research should prioritize evaluation on more challenging and diverse benchmarks that better reflect real-world domain shifts and avoid saturation effects. These include datasets with more severe distribution shifts, larger domain diversity, or deliberately constructed to be dissimilar from common pretraining data.

Bibliography

[1] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang and Chen Change Loy. *Domain Generalization: A Survey*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–20, 2022.

[2] Albert Gu and Tri Dao. *Mamba: Linear-Time Sequence Modeling with Selective State Spaces*, 2024.

[3] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung and James Zou. *Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning*, 2022.

[4] Diederik P. Kingma and Max Welling. *An Introduction to Variational Autoencoders*. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

[5] Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky T. Q. Chen, David Lopez-Paz, Heli Ben-Hamu and Itai Gat. *Flow Matching Guide and Code*, 2024.

[6] Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra and Christopher Ré. *HiPPO: Recurrent Memory with Optimal Polynomial Projections*. *Advances in Neural Information Processing Systems*H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan and H. Lin, editors, vol. 33, pages 1474–1487. Curran Associates, Inc., 2020.

[7] Albert Gu, Karan Goel and Christopher Ré. *Efficiently Modeling Long Sequences with Structured State Spaces*, 2022.

[8] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu and Xinggang Wang. *Vision mamba: efficient visual representation learning with bidirectional state space model*. *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[9] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao and Yunfan Liu. *VMamba: Visual State Space Model*, 2024.

[10] Aaronvan den Oord, Oriol Vinyals and Koray Kavukcuoglu. *Neural Discrete Representation Learning*, 2018.

[11] Martin Arjovsky, Soumith Chintala and Léon Bottou. *Wasserstein generative adversarial networks*. *International conference on machine learning*, pages 214–223. PMLR, 2017.

[12] Tero Karras, Samuli Laine and Timo Aila. *A Style-Based Generator Architecture for Generative Adversarial Networks*, 2019.

[13] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng and Philip S. Yu. *Generalizing to Unseen Domains: A Survey on Domain Generalization*, 2022.

[14] Prasanna Mayilvahanan, Roland S. Zimmermann, Thaddäus Wiedemer, Evgenia Rusak, Attila Juhos, Matthias Bethge and Wieland Brendel. *In Search of Forgotten Domain Generalization. The Thirteenth International Conference on Learning Representations*, 2025.

[15] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March and Victor Lempitsky. *Domain-adversarial training of neural networks. Journal of machine learning research*, 17(59):1–35, 2016.

[16] Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette and Mario Marchand. *Domain-adversarial neural networks. arXiv preprint arXiv:1412.4446*, 2014.

[17] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales and Tao Xiang. *Deep domain-adversarial image generation for domain generalisation. Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pages 13025–13032, 2020.

[18] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh and Gianfranco Doretto. *Unified deep supervised domain adaptation and generalization. Proceedings of the IEEE international conference on computer vision*, pages 5715–5725, 2017.

[19] Krikamol Muandet, David Balduzzi and Bernhard Schölkopf. *Domain generalization via invariant feature representation. International conference on machine learning*, pages 10–18. PMLR, 2013.

[20] Riccardo Volpi and Vittorio Murino. *Addressing model vulnerability to distributional shifts over image transformation sets. Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7980–7989, 2019.

[21] Yichun Shi, Xiang Yu, Kihyuk Sohn, Manmohan Chandraker and Anil K Jain. *Towards universal representation learning for deep face recognition. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6817–6826, 2020.

[22] Kaiyang Zhou, Yongxin Yang, Yu Qiao and Tao Xiang. *Domain generalization with mixstyle. arXiv preprint arXiv:2104.02008*, 2021.

[23] Kaiyang Zhou, Yongxin Yang, Yu Qiao and Tao Xiang. *Mixstyle neural networks for domain generalization and adaptation. International Journal of Computer Vision*, 132(3):822–836, 2024.

[24] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto and Percy Liang. *Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization*. *arXiv preprint arXiv:1911.08731*, 2019.

[25] David Krueger, Ethan Caballero, Joern Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol and Aaron Courville. *Out-of-distribution generalization via risk extrapolation (rex)*. *International conference on machine learning*, pages 5815–5826. PMLR, 2021.

[26] Zeyi Huang, Haohan Wang, Eric P Xing and Dong Huang. *Self-challenging improves cross-domain generalization*. *European conference on computer vision*, pages 124–140. Springer, 2020.

[27] Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine and Chelsea Finn. *Adaptive risk minimization: Learning to adapt to domain shift*. *Advances in Neural Information Processing Systems*, 34:23664–23678, 2021.

[28] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han Cheol Cho, Seunghyun Park, Yunsung Lee and Sungrae Park. *Swad: Domain generalization by seeking flat minima*. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021.

[29] Pengfei Wang, Zhaoxiang Zhang, Zhen Lei and Lei Zhang. *Sharpness-aware gradient matching for domain generalization*. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3769–3778, 2023.

[30] Xufeng Yao, Yang Bai, Xinyun Zhang, Yuechen Zhang, Qi Sun, Ran Chen, Ruiyu Li and Bei Yu. *Pcl: Proxy-based contrastive learning for domain generalization*. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7097–7107, 2022.

[31] Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee and Clayton Scott. *Domain generalization by marginal transfer learning*. *Journal of machine learning research*, 22(2):1–55, 2021.

[32] Zenan Huang, Haobo Wang, Junbo Zhao and Nenggan Zheng. *idag: Invariant dag searching for domain generalization*. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19169–19179, 2023.

[33] Zhaorui Tan, Xi Yang and Kaizhu Huang. *Rethinking multi-domain generalization with a general learning objective*. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23512–23522, 2024.

[34] Maryam Sultana, Muzammal Naseer, Muhammad Haris Khan, Salman Khan and Fahad Shahbaz Khan. *Self-distilled vision transformer for domain generalization*. *Proceedings of the Asian conference on computer vision*, pages 3068–3085, 2022.

[35] Bo Li, Yifei Shen, Jingkang Yang, Yezhen Wang, Jiawei Ren, Tong Che, Jun Zhang and Ziwei Liu. *Sparse mixture-of-experts are domain generalizable learners*. *arXiv preprint arXiv:2206.04046*, 2022.

[36] Shaocong Long, Qianyu Zhou, Xiangtai Li, Xuequan Lu, Chenhao Ying, Yuan Luo, Lizhuang Ma and Shuicheng Yan. *Dgmamba: Domain generalization via generalized state space model*. *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3607–3616, 2024.

[37] Qi Bi, Jingjun Yi, Hao Zheng, Haolan Zhan, Wei Ji, Yawen Huang and Yuexiang Li. *Dgfamba: Learning flow factorized state space for visual domain generalization*. *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pages 1862–1870, 2025.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin. *Attention Is All You Need*, 2023.

[39] Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu and Lei Zhu. *SegMamba: Long-range Sequential Modeling Mamba For 3D Medical Image Segmentation*, 2024.

[40] Jun Ma, Feifei Li and Bo Wang. *U-Mamba: Enhancing Long-range Dependency for Biomedical Image Segmentation*. *arXiv preprint arXiv:2401.04722*, 2024.

[41] Ziyang Wang, Jian Qing Zheng, Yichi Zhang, Ge Cui and Lei Li. *Mamba-UNet: UNet-Like Pure Visual Mamba for Medical Image Segmentation*, 2024.

[42] Nguyen Thi Hue, Le Hai Anh, Vuong Viet Hung, Tran Le Duc Anh, Dao Duc Thinh, Tran Thi Thao and Hoang Si Hong. *ConvMamba: A Data-Efficient Neural Network for Bearing Fault Diagnosis*. *2024 7th International Conference on Green Technology and Sustainable Development (GTSD)*, pages 165–172, 2024.

[43] Ali Hatamizadeh and Jan Kautz. *Mambavision: A hybrid mamba-transformer vision backbone*. *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25261–25270, 2025.

[44] Aaronvan den Oord, Yazhe Li and Oriol Vinyals. *Representation Learning with Contrastive Predictive Coding*, 2019.

[45] Ting Chen, Simon Kornblith, Mohammad Norouzi and Geoffrey Hinton. *A Simple Framework for Contrastive Learning of Visual Representations*, 2020.

[46] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie and Ross Girshick. *Momentum Contrast for Unsupervised Visual Representation Learning*, 2020.

[47] Jean Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos and Michal Valko. *Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning*. *Advances in Neural Information Processing Systems*H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan and H. Lin, editors, vol. 33, pages 21271–21284. Curran Associates, Inc., 2020.

[48] Xinlei Chen and Kaiming He. *Exploring Simple Siamese Representation Learning*. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15750–15758, 2021.

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger and Ilya Sutskever. *Learning Transferable Visual Models From Natural Language Supervision*, 2021.

[50] Tongzhou Wang and Phillip Isola. *Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere*, 2022.

[51] Sedigheh Eslami and Gerardde Melo. *Mitigate the Gap: Improving Cross-Modal Alignment in CLIP*. *The Thirteenth International Conference on Learning Representations*, 2025.

[52] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu and Mark Chen. *Hierarchical Text-Conditional Image Generation with CLIP Latents*, 2022.

[53] Can Yaras, Siyi Chen, Peng Wang and Qing Qu. *Explaining and mitigating the modality gap in contrastive multimodal learning*. *arXiv preprint arXiv:2412.07909*, 2024.

[54] Ian Goodfellow, Yoshua Bengio and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[55] C.M. Bishop. *Pattern recognition and machine learning*, vol. 4. Springer New York, 2006.

[56] Diederik P Kingma and Max Welling. *Auto-Encoding Variational Bayes*, 2022.

[57] Danilo Jimenez Rezende, Shakir Mohamed and Daan Wierstra. *Stochastic Backpropagation and Approximate Inference in Deep Generative Models*, 2014.

[58] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed and Alexander Lerchner. *beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework*. *International Conference on Learning Representations*, 2017.

[59] Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz and Lawrence Carin. *Cyclical Annealing Schedule: A Simple Approach to Mitigating KL Vanishing*, 2019.

[60] Gustav Bredell, Kyriakos Flouris, Krishna Chaitanya, Ertunc Erdil and Ender Konukoglu. *Explicitly Minimizing the Blur Error of Variational Autoencoders*, 2023.

[61] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle and Ole Winther. *Autoencoding beyond pixels using a learned similarity metric*, 2016.

[62] Danilo Jimenez Rezende and Shakir Mohamed. *Variational Inference with Normalizing Flows*, 2016.

[63] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever and Max Welling. *Improved Variational Inference with Inverse Autoregressive Flow*. *Advances in Neural Information Processing Systems*D. Lee, M. Sugiyama, U. Luxburg, I. Guyon and R. Garnett, editors, vol. 29. Curran Associates, Inc., 2016.

[64] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville and Yoshua Bengio. *Generative Adversarial Networks*, 2014.

[65] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin and Aaron Courville. *Improved Training of Wasserstein GANs*, 2017.

[66] Jonathan Ho, Ajay Jain and Pieter Abbeel. *Denoising Diffusion Probabilistic Models*, 2020.

[67] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon and Ben Poole. *Score-Based Generative Modeling through Stochastic Differential Equations*, 2021.

[68] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan and Surya Ganguli. *Deep unsupervised learning using nonequilibrium thermodynamics*. *International conference on machine learning*, pages 2256–2265. pmlr, 2015.

[69] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon and Ben Poole. *Score-based generative modeling through stochastic differential equations*. *arXiv preprint arXiv:2011.13456*, 2020.

[70] Yang Song and Stefano Ermon. *Generative modeling by estimating gradients of the data distribution*. *Advances in neural information processing systems*, 32, 2019.

[71] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li and Jun Zhu. *Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps*. *Advances in neural information processing systems*, 35:5775–5787, 2022.

[72] Tero Karras, Miika Aittala, Timo Aila and Samuli Laine. *Elucidating the design space of diffusion-based generative models*. *Advances in neural information processing systems*, 35:26565–26577, 2022.

[73] Jiaming Song, Chenlin Meng and Stefano Ermon. *Denoising diffusion implicit models*. *arXiv preprint arXiv:2010.02502*, 2020.

[74] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel and Matt Le. *Flow Matching for Generative Modeling*, 2023.

[75] Michael S. Albergo and Eric Vanden-Eijnden. *Building Normalizing Flows with Stochastic Interpolants*, 2023.

[76] Michael S. Albergo, Nicholas M. Boffi and Eric Vanden-Eijnden. *Stochastic Interpolants: A Unifying Framework for Flows and Diffusions*, 2023.

[77] Qihao Liu, Xi Yin, Alan Yuille, Andrew Brown and Mannat Singh. *Flowing from Words to Pixels: A Noise-Free Framework for Cross-Modality Evolution*. *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2755–2765, 2025.

[78] Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf and Yoshua Bengio. *Improving and generalizing flow-based generative models with minibatch optimal transport*, 2024.

[79] Richard Sinkhorn and Paul Knopp. *Concerning nonnegative matrices and doubly stochastic matrices*. *Pacific Journal of Mathematics*, 21:343–348, 1967.

[80] Sara Beery, Grant Van Horn and Pietro Perona. *Recognition in terra incognita*. *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018.

[81] Da Li, Yongxin Yang, Yi Zhe Song and Timothy M. Hospedales. *Deeper, Broader and Artier Domain Generalization*. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[82] Chen Fang, Ye Xu and Daniel N Rockmore. *Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias*. *Proceedings of the IEEE international conference on computer vision*, pages 1657–1664, 2013.

[83] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty and Sethuraman Panchanathan. *Deep hashing network for unsupervised domain adaptation*. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017.

[84] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. *Deep Residual Learning for Image Recognition*, 2015.

[85] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles and Herve Jegou. *Training data-efficient image transformers amp; distillation through attention*. *Proceedings of the 38th International Conference on Machine Learning*Marina Meila and Tong Zhang, editors, vol. 139 in *Proceedings of Machine Learning Research*, pages 10347–10357. PMLR, 2021.

[86] Jia Deng, Wei Dong, Richard Socher, Li Jia Li, Kai Li and Li Fei-Fei. *ImageNet: A large-scale hierarchical image database*. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[87] Jules Sanchez, Jean Emmanuel Deschaud and François Goulette. *Domain generalization of 3d semantic segmentation in autonomous driving*. *Proceedings of the IEEE/CVF international conference on computer vision*, pages 18077–18087, 2023.

[88] Baolu Li, Jinlong Li, Xinyu Liu, Runsheng Xu, Zhengzhong Tu, Jiacheng Guo, Qin Zou, Xiaopeng Li and Hongkai Yu. *V2x-dgw: Domain generalization for multi-agent perception under adverse weather conditions*. *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 974–980. IEEE, 2025.

[89] Jing Du, John Zelek and Jonathan Li. *Weather-aware autopilot: Domain generalization for point cloud semantic segmentation in diverse weather scenarios*. *ISPRS Journal of Photogrammetry and Remote Sensing*, 218:204–219, 2024.

[90] Jee Seok Yoon, Kwanseok Oh, Yooseung Shin, Maciej A Mazurowski and Heung Il Suk. *Domain generalization for medical image analysis: A review*. *Proceedings of the IEEE*, 2024.

[91] Cheng Ouyang, Chen Chen, Surui Li, Zeju Li, Chen Qin, Wenjia Bai and Daniel Rueckert. *Causality-inspired single-source domain generalization for medical image segmentation*. *IEEE Transactions on Medical Imaging*, 42(4):1095–1106, 2022.

[92] Haoliang Li, YuFei Wang, Renjie Wan, Shiqi Wang, Tie Qiang Li and Alex Kot. *Domain generalization for medical imaging classification with linear-dependency regularization*. *Advances in neural information processing systems*, 33:3118–3129, 2020.

[93] Ziwei Niu, Shuyi Ouyang, Shiao Xie, Yen wei Chen and Lanfen Lin. *A survey on domain generalization for medical image analysis*. *arXiv preprint arXiv:2402.05035*, 2024.

[94] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: a modern approach*. Pearson, 4η έκδοση, 2020.

[95] Tom M Mitchell. *Machine learning*. McGraw Hill New York, 1997.

[96] Trevor Hastie, Robert Tibshirani and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2η έκδοση, 2009.

[97] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 2η έκδοση, 2018.

[98] Richard Szeliski. *Computer Vision: Algorithms and Applications*. Springer-Verlag, 1η έκδοση, 2010.

[99] Mei Wang and Weihong Deng. *Deep visual domain adaptation: A survey*. *Neurocomputing*, 312:135–153, 2018.

[100] Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed and Hamid R. Arabnia. *A Brief Review of Domain Adaptation*. *Advances in Data Science and Information Engineering*Robert Stahlbock, Gary M. Weiss, Mahmoud Abou-Nasr, Cheng Ying Yang, Hamid R. Arabnia and Leonidas Deligiannidis, editors, pages 877–894, Cham, 2021. Springer International Publishing.

[101] Gabriela Csurka. *Domain adaptation for visual applications: A comprehensive survey*. *arXiv preprint arXiv:1702.05374*, 2017.

[102] Jian Liang, Ran He and Tieniu Tan. *A comprehensive survey on test-time adaptation under distribution shifts*. *International Journal of Computer Vision*, 133(1):31–64, 2025.

[103] Dian Chen, Dequan Wang, Trevor Darrell and Sayna Ebrahimi. *Contrastive test-time adaptation*. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 295–305, 2022.

[104] Li Fei-Fei, Robert Fergus and Pietro Perona. *One-shot learning of object categories*. *IEEE transactions on pattern analysis and machine intelligence*, 28(4):594–611, 2006.

[105] Yaqing Wang, Quanming Yao, James T Kwok and Lionel M Ni. *Generalizing from a few examples: A survey on few-shot learning*. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

[106] Wei Wang, Vincent W Zheng, Han Yu and Chunyan Miao. *A survey of zero-shot learning: Settings, methods, and applications*. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–37, 2019.

[107] Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, Xi Zhao Wang and Q. M. Jonathan Wu. *A Review of Generalized Zero-Shot Learning Methods*. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4051–4070, 2023.

[108] Sinno Jialin Pan and Qiang Yang. *A survey on transfer learning*. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[109] Karl Weiss, Taghi M Khoshgoftaar and DingDing Wang. *A survey of transfer learning*. *Journal of Big data*, 3(1):9, 2016.

[110] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong and Qing He. *A comprehensive survey on transfer learning*. *Proceedings of the IEEE*, 109(1):43–76, 2020.

[111] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk and Jenia Jitsev. *LAION-5B: An open large-scale dataset for training next generation image-text models*. *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[112] C. Villani. *Optimal Transport: Old and New*. Springer Berlin Heidelberg, 2008.

[113] S. Kullback and R. A. Leibler. *On Information and Sufficiency*. *Ann. Math. Statist.*, 22, 1951.

[114] Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette and Mario Marchand. *Domain-adversarial neural networks*. *arXiv preprint arXiv:1412.4446*, 2014.

[115] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon and Donggeun Yoo. *Reducing domain gap by reducing style bias*. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8690–8699, 2021.

[116] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra and Christopher Ré. *FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness*, 2022.

[117] Diederik P. Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever and Max Welling. *Improving Variational Inference with Inverse Autoregressive Flow*, 2017.

[118] Kihyuk Sohn, Honglak Lee and Xinchen Yan. *Learning Structured Output Representation using Deep Conditional Generative Models*. *Advances in Neural Information Processing Systems*C. Cortes, N. Lawrence, D. Lee, M. Sugiyama and R. Garnett, editors, vol. 28. Curran Associates, Inc., 2015.

[119] Arash Vahdat and Jan Kautz. *NVAE: A Deep Hierarchical Variational Autoencoder*, 2021.

[120] Youssef Kossale, Mohammed Airaj and Aziz Darouichi. *Mode Collapse in Generative Adversarial Networks: An Overview*. *2022 8th International Conference on Optimization and Applications (ICOA)*, pages 1–6, 2022.

[121] Martin Arjovsky and Léon Bottou. *Towards principled methods for training generative adversarial networks*. *arXiv preprint arXiv:1701.04862*, 2017.

[122] Zhitong Ding, Shuqi Jiang and Jingya Zhao. *Take a close look at mode collapse and vanishing gradient in GAN*. *2022 IEEE 2nd International Conference on Electronic Technology, Communication and Information (ICETCI)*, pages 597–602, 2022.

[123] Alec Radford, Luke Metz and Soumith Chintala. *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks*, 2016.

[124] Tero Karras, Timo Aila, Samuli Laine and Jaakko Lehtinen. *Progressive Growing of GANs for Improved Quality, Stability, and Variation*, 2018.