Advancing Self-Supervised Visual Representation Learning: From Domain-Specific Foundations to Interpretable, Disentangled Structure

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Maofeng Tang

August 2026

Acknowledgements

My deepest gratitude goes to my advisor, Dr. Hairong Qi, for her guidance, support, and encouragement throughout my research journey. Over the past few years, she has met with me regularly to review my progress and has offered insightful suggestions whenever I ran into problems. Her meticulous revisions of my manuscripts, sentence by sentence, have greatly improved my writing. I am especially thankful for her support during the times when I struggled to balance research with my responsibilities at home. What she taught me, though, was never only about research. I have watched her extend patience and care to students in difficulty, whether or not they were her own, and in doing so she showed me what it means to be a better person as well as a better researcher. She also works with an energy I still find remarkable, following new directions in the field closely and putting in the hours to keep pace with them. That example has stayed with me. Her expertise and mentorship have been crucial to the completion of this dissertation.

I would also like to thank my committee members, Drs. Xiongye Xiao, Ling Liang, and Hector Santos Villalobos. Their valuable suggestions, professional attitudes, extensive knowledge, and inspiring passion have motivated me to become a better young researcher.

To my dearest wife, Ming, thank you for your unwavering love, patience, and belief in me, even during the most challenging moments of this journey. You are the one who truly knows what it took me to get here over the past few years. Thank you for listening, comforting, and encouraging me. You have been my greatest source of

strength, sharing every joy and overcoming every difficulty alongside me. To my son, Max, you are my priceless treasure. Thank you for coming into my life. Your joy and energy have been a constant source of motivation for me. Finally, to everyone I met along the way during the past few years, thank you. Every experience has been unique, and each moment has contributed to the completion of this dissertation. I am truly grateful for this journey.

Abstract

Self-supervised visual representation learning has advanced rapidly, yet it remains unclear whether representations can be simultaneously high-performing and interpretable, or whether these properties trade off against one another. This dissertation argues that they are compatible, provided the structure a representation needs is built into the architecture and training objective rather than left for scale to discover implicitly. The argument is developed across three phases, each isolating the mechanism responsible for its improvement.

The first phase establishes the argument in remote sensing, where ground sampling distances differ by 10–50× and dense prediction demands both global semantics and pixel-level precision. Global-local balanced contrastive learning (mCL-LC) adds an explicit local-consistency loss alongside global contrastive objectives. Cross-Scale masked autoencoding (MAE) adds explicit cross-scale consistency and reconstruction objectives, so that representations remain related across scale. Fine-tuned on construction-phase classification across Landsat, Sentinel, and WorldView imagery, the latter exceeds a fully supervised baseline (68.5% versus 61.4% overall accuracy) with less labeled data.

The second phase asks whether this principle is specific to remote sensing. Multi-objective masked image modeling combines complementary learning signals but weights them with a single global scalar applied uniformly to every patch. ExPLoRe repurposes Mixture-of-Experts dispatch weights as learned per-patch loss coefficients, coupling each objective's gradient directly to the router. A detach ablation isolates the

mechanism: added expert capacity without that pathway is actively harmful (74.1% versus a 75.7% baseline in k-NN accuracy on ImageNet-1K), whereas the coupled version reaches 76.2%.

The third phase turns to separability. XTRA introduces factor tokens, dedicated learnable tokens constrained by a minimum volume constraint to occupy geometrically independent positions. On ImageNet-1K, disentanglement improves markedly over DINOv2 (SEPIN@1 of 3.95 versus 0.47) while accuracy also improves. On PartImageNet, individual factor tokens correspond to object parts, with the largest gains on articulated parts whose boundaries are hardest to learn implicitly. Removing the constraint reduces XTRA's +4.5 mIoU improvement over DINOv2 to +1.2.

Across all three phases, the mechanism is the same: structure declared in the objective or architecture, not hoped for as a byproduct of scale.

Table of Contents

1 Introduction 1

1.1 The Evolution of Visual Representation Learning 1

1.2 Research Overview and Narrative 2

1.3 Contributions 4

1.4 Dissertation Organization 5

2 Related Work 7

2.1 Preliminaries 7

2.1.1 General Setup 7

2.1.2 Vision Transformers 8

2.1.3 The Contrastive Objective 8

2.1.4 The Masked Modeling Objective 9

2.1.5 Knowledge Distillation and Self-Distillation 9

2.1.6 Evaluation Protocols and Metrics 10

2.2 Visual Representation in Remote Sensing 10

2.3 Contrastive Learning 11

2.4 Multi-Scale Representation Learning 12

2.5 Masked Image Modeling (MIM) 14

2.6 Disentanglement and Latent Structure 15

3 Foundations in Remote Sensing: Solving Scale and Consistency Challenges 18

3.1 The Remote Sensing Challenge . . . 19
3.2 Global-Local Balanced Contrastive Learning . . . 20
3.2.1 Motivation: The Local Consistency Problem . . . 21
3.2.2 Framework Design . . . 22
3.2.3 Experimental Validation . . . 27
3.2.4 Discussion . . . 28
3.3 Cross-Scale MAE: Multi-Scale Masked Image Modeling . . . 30
3.3.1 Motivation: Multi-Scale Challenges in Masked Image Modeling 30
3.3.2 Cross-Scale MAE Architecture . . . 31
3.3.3 Experimental Results . . . 37
3.3.4 Fine-Tuning Validation: The SMART Dataset . . . 44
3.3.5 Discussion . . . 44
3.4 Summary . . . 45

4 Generalizing Consistency: Adaptive Loss Routing for Multi-Objective Masked Image Modeling 46
4.1 Generalizing the Consistency Principle . . . 47
4.2 Patch-Level Loss Routing for Multi-Objective Masked Image Modeling 48
4.2.1 The Patch-Level Heterogeneity Problem . . . 48
4.2.2 Repurposing Expert Routing as Per-Patch Loss Coefficients . 49
4.2.3 Mechanism Validation and Expert Scaling . . . 51
4.2.4 Downstream Transfer and Prior Work . . . 52
4.3 Summary . . . 53

5 Refinement through Disentanglement 56
5.1 The Interpretability Gap . . . 57
5.2 XTRA . . . 58
5.2.1 XTRA Architecture . . . 58
5.2.2 Minimum Volume Constraint . . . 60
5.2.3 Training Objective and Methodology . . . 63

5.2.4 Experimental Validation 65
5.2.5 Factor Token Visualization 69
5.3 Summary 70

6 Discussion and Future Directions 73
6.1 Limitations of the Completed Work 73
6.2 Controllable Generation via Disentangled Latent Spaces 75
6.3 Efficient, Multi-Scale Remote Sensing Foundation Models 76
6.4 Summary 77

7 Conclusion 78
7.1 Summary of Contributions 78
7.2 The Case for Explicit Structure 80
7.3 Closing Remarks 81

Bibliography 82

Publications 96

Vita 99

List of Tables

3.1 COMPARISON WITH STATE-OF-THE-ART CONTRASTIVE LEARNING FRAMEWORKS IN TERMS OF PIXEL-BASED SEMANTIC SEGMENTATION ACCURACY. NOTE THAT SUPERVISED BASELINE REFERS TO DEEPLAB V3+ 27

3.2 ZERO-SHOT GENERALIZATION COMPARISON IN SITES TRANSFER (OA) 28

3.3 The information of datasets used in Sec. 3.3.3 37

3.4 Average KNN accuracy with different scale ratios (12.5%, 25%, 50%, 100%) 40

3.5 Linear classification fine-tuning performance on fMoW-RGB 40

3.6 Semantic segmentation performance on Potsdam & Vaihingen (mIoU) 40

4.1 Mechanism validation on the Token+CLS configuration (ImageNet-1K, k-NN@20). Repurposing dispatch weights as loss coefficients, with gradients coupled to the router, is what turns additional expert capacity into a benefit rather than a liability. 51

4.2 Comparison with published self-supervised methods on ImageNet-1K (ViT-Base encoders). Lin. = linear-probe top-1, FT = fine-tuning top-1. 53

5.1 Representation disentanglement score with SEPIN@k on ImageNet-1k, where k denotes the top-k dimensions (higher is better). 66

5.2 Part segmentation on PartImageNet. XTRA achieves superior part-level mIoU, with largest improvements on articulated parts. 67
5.3 Per-part breakdown for the "Quadruped" category. 67
5.4 Evaluation of representation from pre-trained model in KNN and Linear Probing (%). 68
5.5 Evaluation of representation without pre-trained teacher network in KNN and Linear Probing on IN-1K(%). 68
5.6 Evaluation of representation from pre-trained model with different downstream tasks (%). 69

List of Figures

3.1 Illustration of the mCL-LC architecture. 23

3.2 ROIs in negative pair. The red, yellow, blue boxes are the matching ROIs in a negative image pair, but the matching ROIs bounded by the red and yellow boxes are actually similar. 25

3.3 Class-wise F1-score on Nice and Nantes Saint. 29

3.4 The architecture of Cross-Scale MAE comprises an encoder and a decoder (a). The encoder (b-top) employs a vision transformer (ViT) backbone, specifically a ViT-base with 12 self-attention blocks. The decoder (b-bottom) uses a lightweight ViT backbone with 8 self-attention blocks. A single satellite image undergoes scale augmentation through random cropping and resizing. The contrastive loss is computed using encoder outputs for the two scale inputs. The cross-prediction loss (in MSE) is applied to the decoder output of the last self-attention block. The reconstruction loss compares the predicted masked patch with the actual input. 33

3.5 The KNN classification of Cross-Scale MAE for different datasets. . . 39

3.6 Reconstruction performance on the CoCo Dataset 41

3.7 Multiscale reconstruction performance on the CoCo dataset 42

3.8 Zero-shot reconstruction in fMoW with Cross-Scale MAE pre-trained on CoCo 43

4.1 ExPLoRe framework overview. Soft Mixture-of-Experts (Soft-MoE) is integrated into the student encoder for patch-level adaptive loss weighting. The student encoder (ViT-Base with alternating MoE blocks) processes patches while a frozen CLIP teacher provides semantic targets. Soft-MoE dispatch weights serve as per-patch loss coefficients, with each expert weighting a different training objective. Loss-coupling, in which loss gradients flow through the dispatch weights back to the router, is the mechanism that enables learned, content-dependent specialization. 55

5.1 Illustration of XTRA built upon the dual-stream self-knowledge distillation network. Top: teacher network. Bottom: The multi-stage aggregation student network. 59

5.2 Illustration of a 2-stage aggregation of factor tokens 60

5.3 (a) A geometric illustration of the two loss terms within the latent loss (Eq. 5.4) where the minimum volume constraint, $J(F)$, serves as the internal force pointing inward and the patch reconstruction constraint, $\mathcal{L}_{factor}$, serves as the external force pointing outward. (b) Illustration of how the factor tokens and patch tokens evolve across two stages of aggregations. 61

5.4 XTRA enables disentangled attention maps pertaining to consistent parts across multiple objects in the scene. 70

5.5 Visualization of Factor Tokens 71

5.6 Visualization of Factor Tokens (Other Objects) 72

Nomenclature

Sets and Spaces

$\mathcal{D}$	Unlabeled image collection used for pretraining
$\mathcal{X}$	Input image space
$\mathcal{Z}$	Latent representation space
$\mathcal{M}$	Index set of masked patches
$\mathcal{V}$	Index set of visible (unmasked) patches
Λ^-	Set of negative samples in a mini-batch
Δ^{K-1}	Probability simplex on K vertices
$\mathbb{R}^D$	D-dimensional Euclidean space

Dimensions and Counts

N	Number of patches in a sequence; also the number of samples in a mini-batch, where indicated
D	Model (embedding) dimension of a token
K	Number of factor tokens
B	Batch size

L	Number of transformer layers; also the number of aggregation stages in XTRA
H, W, C	Image height, width, and channel count
n	Patch side length in pixels
N_R	Number of regions of interest sampled per patch

Tokens and Matrices

$\mathbf{p}_i$	Patch token i
$\mathbf{f}_k$	Factor token k
$\mathbf{F}$	Factor token matrix, $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_K] \in \mathbb{R}^{D \times K}$
$\mathbf{G}$	Gram matrix of the factor tokens, $\mathbf{G} = \mathbf{F}^T\mathbf{F}$
$\mathbf{w}_i$	Mixing weights expressing patch i as a combination of factors
$\mathbf{I}$	Identity matrix
z	Latent representation of an image
$\mathbf{t}$	Teacher feature target
$\mathbf{g}_i$	Learnable aggregation token i

Networks and Maps

f_θ	Encoder with parameters θ
g_ϕ	Decoder with parameters ϕ
$g_f(\cdot)$	Feature-level projection head
$g_s(\cdot)$	Semantic-level projection head

$g_p(\cdot)$	Cross-scale prediction head
T, S	Teacher and student networks in self-distillation
$\mathrm{sim}(\cdot, \cdot)$	Cosine similarity
$\mathcal{I}(\cdot\,;\cdot)$	Mutual information between two variables

Losses and Objectives

$\mathcal{L}_{\text{total}}$	Complete training objective
ℓ_{InfoNCE}	InfoNCE contrastive loss
ℓ_{NTX}	NT-Xent contrastive loss
$\mathcal{L}_{fl}$	Feature-level contrastive loss (mCL-LC)
$\mathcal{L}_{sl}$	Semantic-level contrastive loss (mCL-LC)
$\mathcal{L}_{lc}$	Local consistency loss (mCL-LC)
$\mathcal{L}_{cc}$	Cross-scale consistency loss at the encoder (Cross-Scale MAE)
$\mathcal{L}_{cp}$	Cross-scale prediction loss at the decoder (Cross-Scale MAE)
$\mathcal{L}_{re}$	Reconstruction loss (Cross-Scale MAE)
$\mathcal{L}_{\text{token}}$	Token-level distillation loss (ExPLoRe)
$\mathcal{L}_{\text{cls}}$	Class-token alignment loss (ExPLoRe)
$\mathcal{L}_{\text{pixel}}$	Pixel reconstruction loss (ExPLoRe)
$\mathcal{L}_{\text{entropy}}$	Per-expert entropy regularizer (ExPLoRe)
$\mathcal{L}_{\text{distill}}$	Self-distillation loss (XTRA)
$\mathcal{L}_{\text{factor}}$	Patch reconstruction loss from factor tokens (XTRA)

$\mathcal{L}_{\text{latent}}$	Combined latent objective, $\mathcal{L}_{\text{factor}}$ plus the volume penalty (XTRA)
ℓ_{huber}	Huber loss

Hyperparameters

λ_{distill}	Weight on the distillation loss (XTRA)
λ_{factor}	Weight on the factor reconstruction loss (XTRA)
λ_{volume}	Weight on the minimum volume penalty (XTRA)
λ_e	Per-expert entropy regularization weight (ExPLoRe)
τ	Temperature parameter in contrastive objectives
m	Masking ratio
r_h, r_l	Higher and lower scale ratios in multi-scale augmentation
g_h, g_l	Ground sampling distances of the two augmented views
σ	Noise bound in the linear mixing model

Mixture-of-Experts Notation

E	Number of experts in a Soft-MoE layer
$\mathbf{D}$	Dispatch weights, normalized over patches for each expert
$\mathbf{C}$	Combine weights, normalized over experts for each patch
$\mathbf{\Phi}$	Learnable expert parameters used to compute routing logits
$\mathbf{L}$	Routing logits
s	Learned scale applied to the routing logits
H_e	Entropy of expert e's dispatch distribution across patches

Abbreviations

CLS	Class token
DCI	Disentanglement, Completeness, Informativeness
GSD	Ground sampling distance
MAE	Masked autoencoder
$mIoU$	Mean intersection over union
MIM	Masked image modeling
MoE	Mixture of experts
MVC	Minimum volume constraint
SSL	Self-supervised learning
ViT	Vision transformer

Chapter 1

Introduction

1.1 The Evolution of Visual Representation Learning

Deep learning changed computer vision by replacing hand-crafted features with learned ones Krizhevsky et al. (2012); He et al. (2016); Lowe (2004). That change came at a cost. Its early successes in image classification depended on large annotated datasets, ImageNet Deng et al. (2009) above all, and annotation does not scale with the volume of imagery now being collected.

Self-supervised learning removes that dependency Jing and Tian (2020); Liu et al. (2021). Rather than labels, it derives its supervisory signal from the data itself through a pretext task, and the representations it produces now match or exceed supervised ones on downstream benchmarks while requiring no manual annotation during pretraining at all, which is why much of the last five years of representation learning has been a question of which pretext task to choose Chen et al. (2020a); He et al. (2020a).

Two families dominate. Contrastive methods Chen et al. (2020a); He et al. (2020a); Grill et al. (2020) maximize agreement between augmented views of the same image and are strong at global semantics. Masked image modeling (MIM) instead

removes part of the input and asks the model to restore it: the masked autoencoder (MAE) He et al. (2022b) reconstructs raw pixels, BEiT Bao et al. (2021) predicts discrete visual tokens, and MaskDistill Wei et al. (2022) predicts the features of a pretrained teacher such as CLIP Radford et al. (2021). Both families work. Neither asks what kind of content it is being applied to.

We take that last point as the starting condition for this dissertation. A contrastive objective imposes the same invariance on every image, and a masking objective asks the same reconstruction of every patch, whatever that image or patch happens to contain. For natural images, the assumption is usually harmless. For satellite imagery it is not: a fixed patch spans tens of meters or tenths of a meter depending on the sensor Yuan et al. (2021), and treating those two cases identically discards the one thing that distinguishes them. We read this as evidence for a more general claim. When content is heterogeneous in a way that matters, the structure needed to handle it should be declared in the architecture or the objective, not left for scale and capacity to discover.

There is a second half to the argument, and it does not follow from the first. A representation can encode scale explicitly, and route learning signals per patch, and still be unreadable: no individual factor can be examined without the others moving with it Bengio et al. (2013). Declaring that structure exists is not the same as making its parts separable. We argue both halves, and the second is what Chapter 5 is for.

1.2 Research Overview and Narrative

We present three contributions. They are numbered as phases because that is the order in which we did the work, not because each one waits on the one before it: Phases 1 and 3 make the two halves of the argument, and Phase 2 tests whether the first half holds outside the domain it came from.

Phase 1: Foundations in Remote Sensing: Solving Scale and Consistency Challenges Remote sensing imagery presents unique challenges that

make it an ideal testbed for advancing representation learning. Satellite images exhibit extreme scale variations due to differences in ground sampling distance, capturing objects ranging from individual vehicles to entire cities within similar-sized crops Cheng et al. (2017b). This stage addresses how self-supervised learning can be adapted for such domain-specific challenges through two contributions: a global-local balanced contrastive learning framework Tang et al. (2023b) that enforces local consistency while maintaining global semantic understanding, and Cross-Scale MAE Tang et al. (2023a), a masked image modeling framework designed for multi-scale scenarios. Building on MAE He et al. (2022b) and inspired by domain-specific adaptations like SatMAE Cong et al. (2022) and Scale-MAE Reed et al. (2022), Cross-Scale MAE achieved state-of-the-art results across multiple remote sensing benchmarks Georgiou et al. (2024).

Phase 2: Generalizing Consistency: Adaptive Loss Routing The success of domain-specific solutions raised a critical question: does the broader principle behind Phase 1—that explicit mechanisms for handling spatial heterogeneity outperform uniform treatment—generalize beyond remote sensing's particular form of heterogeneity, scale? Through collaborative work (equal contribution), this stage identified an analogous heterogeneity problem in multi-objective masked image modeling: existing methods weight complementary learning signals (token distillation, CLS alignment, pixel reconstruction) with a single global scalar per objective, applied uniformly across every patch. The resulting framework, ExPLoRe Georgiou et al. (2026), repurposes Mixture-of-Experts routing to learn per-patch loss weights instead, improving representation quality on ImageNet Deng et al. (2009) while using fewer inference FLOPs than comparably sized baselines. This work confirmed that the Phase 1 principle generalizes beyond scale variation specifically, but also revealed a crucial limitation: high performance does not guarantee interpretable structure.

Phase 3: Refinement through Disentanglement: From Performance to Interpretable Structure High performance on benchmarks is insufficient for applications requiring controllable generation, compositional understanding, and

human interpretability Bengio et al. (2013); Higgins et al. (2018). Strong benchmarks say nothing about whether a representation can be read. We turned in Phase 3 to whether its semantic factors can be separated at all. Building on recent work adding extra tokens to vision transformers for other purposes Darcet et al. (2023); Xu et al. (2022); Seitzer et al. (2023), we introduced dedicated *factor tokens*, each given its own place in the sequence to hold a distinct semantic factor Tang and Qi (2026). What keeps them distinct is a Minimum Volume Constraint (MVC), a geometric penalty on redundancy between factors: correlated factors span less of the representation space than distinct ones, so penalizing that collapse is what holds them apart. XTRA improves disentanglement Eastwood and Williams (2018); Locatello et al. (2019) without giving up downstream accuracy. The two are not in tension.

Together, these three phases trace a single arc: from domain-specific foundations that establish the value of explicit, structured mechanisms over uniform treatment (Phase 1), to a demonstration that this same principle generalizes to natural images (Phase 2), to a demonstration that a comparable commitment to explicit structure yields not just better performance but genuinely interpretable representations (Phase 3). Chapter 6 returns to the possibilities this progression opens up, including controllable generation and efficient multi-scale foundation models for remote sensing, as directions for future work.

1.3 Contributions

This dissertation makes three contributions, grouped around two claims. Chapters 3 and 5 carry the argument in its two halves: that structure should be declared rather than left to emerge, and that declared structure is useful only when its factors are separable. Chapter 4 stands between them and tests whether the first half is an artifact of remote sensing. It works on a third axis of heterogeneity, which is the learning signal a patch requires rather than the scale at which it was captured or the

semantic factor it carries, and is therefore a parallel test of the same claim rather than a step from one to the other.

Domain-Specific Foundations: Global-local balanced contrastive learning framework Tang et al. (2023b) showing that local consistency constraints improve dense prediction in aerial imagery. Cross-Scale MAE Tang et al. (2023a) provides a multi-scale masking and reconstruction framework achieving state-of-the-art results across remote sensing benchmarks, with validated fine-tuning strategies Georgiou et al. (2024).

Adaptive Loss Routing: ExPLoRe Georgiou et al. (2026) (equal contribution), which repurposes Mixture-of-Experts dispatch weights as learned, per-patch loss coefficients for multi-objective masked image modeling, demonstrating that this mechanism, rather than expert capacity alone, drives improved representation quality on ImageNet, with dedicated downstream-transfer recipes closing the adaptation gap on ADE20K segmentation.

Disentangled Representations: XTRA Tang and Qi (2026) introduces factor tokens in vision transformers designed to capture distinct semantic factors. The Minimum Volume Constraint (MVC) provides a geometrically motivated objective encouraging statistical independence. The framework achieves strong disentanglement metrics while maintaining downstream performance, with visualization tools for understanding the learned structure.

1.4 Dissertation Organization

The remainder of this dissertation is organized as follows. The three contributions are presented in Chapters 3, 4, and 5, and are referred to as Phases 1 through 3 where the order in which the work was carried out is relevant. **Chapter 2** reviews the preliminaries used throughout, followed by background on representation learning in remote sensing, contrastive learning, multi-scale representation learning, masked image modeling, and disentanglement. **Chapter 3** presents the Phase 1 foundations

in remote sensing, covering global-local balanced contrastive learning and Cross-Scale MAE. **Chapter 4** presents Phase 2, extending the consistency principle to patch-level loss routing for multi-objective masked image modeling. **Chapter 5** presents Phase 3, refinement through disentangled representations via factor tokens and the Minimum Volume Constraint. **Chapter 6** opens with the limitations of the completed work, drawn from the publications' own discussions, and derives from them the two directions those limitations point toward: controllable generation from disentangled latent spaces, and efficient multi-scale foundation models for remote sensing. Neither is offered as a research plan. **Chapter 7** concludes by synthesizing the three contributions and setting out the case for explicit structure that runs through all of them.

Chapter 2

Related Work

2.1 Preliminaries

This section fixes the notation and the standard objectives used throughout the dissertation. Readers familiar with self-supervised learning may skip to Section 2.2; the definitions are collected here so that later chapters can refer to them rather than restate them.

2.1.1 General Setup

Let $\mathcal{D} = \{x_i\}_{i=1}^{N}$ be a large unlabeled image collection, with $x_i \in \mathbb{R}^{H \times W \times C}$. The aim of self-supervised representation learning is to obtain an encoder $f_\theta : \mathcal{X} \to \mathcal{Z}$ mapping an image x to a latent representation $z \in \mathbb{R}^d$ that carries semantic information transferable to downstream tasks, without using labels during pretraining. The parameters θ are learned by optimizing a pretext objective, and the quality of f_θ is judged by how well z supports tasks it was never trained on.

2.1.2 Vision Transformers

The Vision Transformer (ViT) (Dosovitskiy et al., 2020) partitions an image into a grid of non-overlapping square patches, projects each patch linearly into a D-dimensional embedding, and processes the resulting sequence with a standard transformer encoder. Each element of that sequence is a *token*. Tokens corresponding to image regions are *patch tokens*, written $\{\mathbf{p}_i\}_{i=1}^{N}$ where N is the number of patches.

Alongside the patch tokens, a single learnable *class token*, written `[CLS]`, is prepended to the sequence. It corresponds to no image region and is instead free to aggregate information from all of them through self-attention. Its output is conventionally taken as the image-level representation z. The distinction matters for Chapter 5: the `[CLS]` token is a single slot asked to summarize everything, and the factor tokens introduced there are additional slots of the same kind, each asked to summarize something narrower.

2.1.3 The Contrastive Objective

Contrastive learning shapes the embedding space directly, pulling semantically similar views together and pushing dissimilar ones apart. Let $\mathcal{T}$ be a set of stochastic augmentations. For an image x, two views $v = t(x)$ and $v' = t'(x)$ are drawn with $t, t' \sim \mathcal{T}$ and encoded as $z = f_\theta(v)$ and $z' = f_\theta(v')$. The pair (z, z') is a *positive pair*; embeddings of other images form the *negative* set $\mathcal{K}^-$. The standard objective is the InfoNCE loss:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(\text{sim}(z, z')/\tau)}{\exp(\text{sim}(z, z')/\tau) + \sum_{k \in \mathcal{K}^-} \exp(\text{sim}(z, z_k^-)/\tau)} \tag{2.1}$$

where $\text{sim}(\cdot, \cdot)$ is cosine similarity and τ is a temperature. The choice of $\mathcal{T}$ determines what the representation is asked to be invariant to, which is why augmentation design carries so much weight in this family of methods.

2.1.4 The Masked Modeling Objective

Masked image modeling (MIM) takes the opposite approach: rather than specifying invariances, it requires the representation to retain enough information to reconstruct what was removed. The image is tokenized into patches $x_p = \{p_1, \ldots, p_L\}$, a subset $\mathcal{M}$ is masked according to some strategy, and the encoder sees only the visible remainder, $z_{vis} = f_\theta(x_{vis})$. A decoder g_ϕ then reconstructs the masked content, typically under a mean squared error objective:

$$\mathcal{L}_{\text{MIM}} = \sum_{i \in \mathcal{M}} \|p_i - g_\phi(z_{vis})_i\|_2^2 \tag{2.2}$$

The Masked Autoencoder (MAE) (He et al., 2022b) is the canonical instance, and the one referred to throughout this dissertation. It masks a large fraction of patches, encodes only the visible ones, and reconstructs raw pixels with a lightweight decoder that is discarded after pretraining. Two properties matter later: the masking is uniform, so every patch is treated as equally worth reconstructing, and the reconstruction target is fixed, so every patch is asked for the same kind of information. Chapter 4 takes issue with both.

2.1.5 Knowledge Distillation and Self-Distillation

Knowledge distillation trains a *student* network to match the outputs of a *teacher* network rather than ground-truth labels. In the self-supervised setting, the teacher is generally not a separate pretrained model but a copy of the student maintained as an exponential moving average of its weights, an arrangement known as *self-distillation*. The student sees one view of an image and is trained to predict the teacher's output on another, so the supervisory signal is generated by the model itself. DINO (Caron et al., 2021) established this pattern for vision transformers, and it is the training arrangement used in Chapters 4 and 5.

2.1.6 Evaluation Protocols and Metrics

Because pretraining produces no task predictions, representation quality is measured indirectly. Two protocols recur. Under *linear probing*, the encoder is frozen and a single linear classifier is fitted on top, so that accuracy reflects how linearly separable the classes already are in $\mathcal{Z}$. Under *k-nearest-neighbor (k-NN) evaluation*, no parameters are fitted at all: a query is labeled by a vote among its k nearest training embeddings. The k-NN protocol is the stricter of the two, since it cannot compensate for a poorly arranged space.

Task-level results use standard measures. *Overall accuracy* (OA) is the fraction of correctly classified samples. *Mean intersection over union* (mIoU) averages, across classes, the ratio of the intersection to the union of predicted and ground-truth regions, and is the standard measure for semantic and part segmentation. Remote sensing results additionally report *Cohen's kappa*, which corrects agreement for what chance alone would produce.

Disentanglement requires its own measures, since a representation can be accurate and still be entangled. This dissertation reports SEPIN@k (*separateness and informativeness* at k) (Do and Tran, 2019), which quantifies the extent to which the k most informative latent dimensions carry non-redundant information about the data; higher values indicate that individual dimensions capture distinct content rather than repeating one another. It is computed without ground-truth factor labels, which is what makes it applicable to natural images.

2.2 Visual Representation in Remote Sensing

Deep learning in Remote Sensing initially relied on Transfer Learning from ImageNet-pretrained CNNs. However, the domain gap (overhead view, rotational invariance, multispectral data) limits the efficacy of this transfer. Recent works have attempted to adapt SSL to RS. CL-based representation learning methods have been developed,

such as SeCo Manas et al. (2021) and mCL-LC Tang et al. (2023b). SeCo, short for Seasonal Contrast, offers a novel approach to the challenge of insufficient labeled data in remote sensing by leveraging self-supervised learning mechanisms. The mCL-LC method jointly learns global and local representations to improve segmentation performance on satellite images. On the other hand, the generative approach, e.g., MAE, requires the representation to preserve as much of the input information as possible. MAE is an autoencoder with masked prediction, i.e., predicting a property of masked input from unmasked input. Building on the success of masked language models Radford et al. (2018, 2019); Brown et al. (2020); Devlin et al. (2018), a similar mechanism has been investigated in vision tasks and has dramatically advanced the field of SSL Dosovitskiy et al. (2020); He et al. (2022b). Following trends in computer vision, the masked image model has also shown strength in remote sensing. For example, the GFM model Mendieta et al. (2023) excels in representation learning through continual learning, aiming to enhance the applicability of large language models to satellite imagery via knowledge distillation.

Differentiation from Our Work: Most RS-specific SSL methods focus solely on data augmentation (e.g., creating specific views). Our work goes deeper by modifying the *architectural priors* of the learning process itself—specifically through explicit multi-scale consistency objectives and global-local loss balancing—to handle the intrinsic geometric properties of satellite imagery.

2.3 Contrastive Learning

Given the objective in Eq. 2.1, what separates methods in this family is largely how they supply negatives and how they keep the representation from collapsing. SimCLR Chen et al. (2020c) draws negatives from the other images in the mini-batch, which makes performance strongly dependent on batch size. MoCo v2 He et al. (2020b); Chen et al. (2020d) decouples the two by holding negatives in a dynamic queue that can far exceed the batch, and updates the encoder producing them by

momentum so that queued representations do not drift out of step with the encoder being trained. GLCNet Li et al. (2022b) adds a local feature contrastive module alongside the global one, targeting semantic segmentation of high-resolution satellite imagery. Specifically, the SimCLR method Chen et al. (2020c) is based on the idea of instance-wise contrastive learning, which learns by forcing positive samples from the same sample to be similar and negative samples from different samples in a mini-batch to be dissimilar. MoCo v2 He et al. (2020b); Chen et al. (2020d) is also based on the idea of instance-wise contrastive learning, which focuses on obtaining negative samples far beyond the batch size. Therefore, a dynamic queue is proposed to store the features of negative samples, and a momentum-updated encoder is proposed to avoid the consistency problem in the representations of negative samples caused by rapid encoder changes. GLCNet Li et al. (2022b) is a contrastive learning framework based on the global and local features. By adding a local feature contrastive learning module, GLCNet aims for high-resolution satellite image semantic segmentation.

Differentiation from Our Work: While standard CL methods excel at object-centric natural images, they assume global semantic consistency—i.e., that any two crops of an image share the same semantic label. In Remote Sensing (RS), this assumption fails due to high semantic density. Our work (WACV 2023) differs in that it introduces a *local consistency branch* that balances global instance discrimination with pixel-level local constraints, thereby preventing the suppression of fine-grained details.

2.4 Multi-Scale Representation Learning

The multi-scale phenomenon is common in vision tasks. To leverage the information in multi-scale sources, the vision community has proposed to extract multi-scale features using traditional approaches, including, for example, spatial pyramids Rosenfeld and Thurston (1971); Burt and Adelson (1987); Lazebnik et al. (2006), dense sampling of windows Ji et al. (2013); Yan et al. (2012, 2014), and the

combination of them Felzenszwalb et al. (2008). In the last decade, deep convolution neural networks (CNNs) have emerged as the de facto architecture for a wide range of vision tasks. Because of the pooling and multi-convolution kernel operation, the CNN constructs the feature pyramids inherently so that it has been used to build deep multi-scale representations Jiao et al. (2021). CNN-based multi-scale representation learning is commonly approached in two ways. The first approach involves utilizing external or preset factors, which can include multi-scale kernel architecture Szegedy et al. (2015); Kovashka and Grauman (2010); Reininghaus et al. (2015); Zou and Cheng (2021) and multi-scale inputs architecture Dollár and Zitnick (2014); Farabet et al. (2012); Lin et al. (2016). The second approach involves designing internal layers of the network, such as skip and dense connections Singh and Davis (2018); Pohlen et al. (2017); Chen et al. (2017). Over the past three years, there has been a notable surge of interest in applying transformer-based architectures to computer vision tasks. Among these architectures, the Vision Transformer (ViT) Dosovitskiy et al. (2020) stands out as a particularly successful example, in which, compared to CNNs, ViT better balances global and local features. Recently, other efforts have been made toward multi-scale feature learning, including the Multi-scale Deformable Attention and Multi-level Feature Aggregation (MSDAM and MLFAM) network Dong et al. (2022) and the Shunted Self-Attention (SSA) network Ren et al. (2022). These methods have shown strength in multi-level feature extraction for the natural image containing multiple objects of different sizes. However, in remote sensing images, the multi-scale phenomenon is even more prevalent and dynamic. As a result, there has been a growing recognition of the need for customized approaches to address this. The recently published Scale-MAE Reed et al. (2022) is one such example.

Differentiation from Our Work: Essentially, compared with the most recent two state-of-the-art MAE-based satellite imagery representation learning methods, i.e., SatMAE Cong et al. (2022), and Scale-MAE Reed et al. (2022), SatMAE is the first paper that applies MAE to extract representations from satellite images with single and fixed scale. Scale-MAE and our Cross-Scale MAE are both based on MAE

but focus on multi-scale characteristics. Specifically, Scale-MAE modifies the GSD position encoding and applies multi-deconvolution after the decoder to reconstruct images at different scales. Nonetheless, Scale-MAE integrates scale information into the network via hard-coding of a known GSD, and the deconvolution can only yield a specific scale ratio. Our work Cross-Scale MAE (NeurIPS 2023) designs the network to learn the information across different scales. With scale augmentation and multi-level contrastive loss between the scale pair and masked patches reconstruction, Cross-Scale MAE can learn informative and consistent representation across different scales without the need of known GSD. Additionally, we leverage xFormers on a single GPU for pre-training efficiency.

2.5 Masked Image Modeling (MIM)

Masking strategies are crucial to MIM performance, as they determine which parts of the image the model must reconstruct. Within the framework of Section 2.1.4, methods differ chiefly in two choices: what the model is asked to predict, and which patches it is asked to predict. On the first, MAE He et al. (2022b) reconstructs raw pixels, while BEiT Bao et al. (2021) predicts discrete visual tokens produced by a pretrained tokenizer, and iBOT Zhou et al. (2021) follows the BYOL Grill et al. (2020) paradigm in predicting the output of a self-distilled teacher rather than anything in the image itself. The second choice, which patches to mask, is where most subsequent work has concentrated.

More recent works have focused on developing more sophisticated masking techniques other than random and block masking to capture meaningful semantic information. For example, SimMIM Xie et al. (2022) adopts an approach of randomly masking large square patches. MILAN Hou et al. (2022) and AttnMask Kakogeorgiou et al. (2022) utilize the encoder's self-attention mechanisms to produce semantically aware masks. ADIOS Shi et al. (2022) introduces an adversarial masking subnet that learns to mask the most informative patches. SemMAE Li et al. (2022a) incorporates

higher-level semantics into the masking process by using a semantic segmentation model to guide the masking, ensuring that the model focuses on reconstructing semantically significant regions. Evolved Part Masking Feng and Zhang (2023) proposed creating adaptive mask patterns that evolve during training. It starts with simple grid masks and gradually incorporates more complex patterns based on the model's learned attention, guided by an EM algorithm. This allows the masking strategy to adapt to the model's improving understanding of visual structures.

Differentiation from Our Work: Existing MIM methods typically employ random masking strategies designed for redundancy-heavy natural images. In RS, where scale varies drastically, random masking often destroys semantic context or leaves it too easy to reconstruct. Our Cross-Scale MAE (NeurIPS 2023) differs in that it explicitly incorporates multi-scale inputs and cross-scale interactions. Furthermore, existing multi-objective MIM methods weight their combined learning signals with a single global scalar per objective, applied uniformly across every patch; our ExPLoRe work (ECCV 2026) instead repurposes Mixture-of-Experts routing to learn per-patch loss weights, allowing different regions of an image to be weighted according to which learning signal serves them best.

2.6 Disentanglement and Latent Structure

The idea that a transformer should carry tokens which correspond to no part of the input predates vision. BERT (Devlin et al., 2019) introduced the `[CLS]` and `[MASK]` tokens for exactly this purpose, and Visual Prompt Tuning and its variants (Jia et al., 2022; Yoo et al., 2023; Wang et al., 2024a) extended it by injecting small sets of learnable tokens at every layer to adapt a frozen backbone. A parallel line treats tokens as something to remove rather than add: A-ViT (Yin et al., 2022) learns per-token halting probabilities to discard low-value tokens, and Attentive Tokens (Long et al., 2022) merge tokens by learned importance.

The lineage most directly relevant to Chapter 5 runs through the DINO family. DINO (Caron et al., 2021) showed that self-distilled vision transformers develop attention maps that segment objects without ever being trained to, an early and widely cited demonstration that useful structure can emerge from scale and a good objective alone. DINOv2 (Oquab et al., 2024) scaled the recipe and improved the representations, but also made visible a defect the smaller models had hidden: a handful of patch tokens acquire anomalously high norms and act as global scratch space, corrupting the attention maps that made the original result compelling. Darcet et al. (2023) diagnosed this and proposed the fix that gives the phenomenon its name. Their *register tokens* are additional learnable tokens appended to the sequence, corresponding to no image patch, which the model may use for whatever global bookkeeping it was previously doing inside the patch tokens. Adding them restores clean attention maps.

What is notable about that progression is where the structure comes from at each step. In DINO, the segmentation behavior is emergent. In DINOv2, the artifacts are emergent. In the register work, the remedy is explicit: rather than hoping the model would stop misusing patch tokens, the architecture is given somewhere else to put the information. The registers are not, however, asked to hold anything in particular; they are unconstrained slots, and what accumulates in them is again left to emerge. Chapter 5 takes the remaining step, giving the extra tokens both a designated purpose and a constraint that keeps them distinct.

None of the disentanglement literature surveyed below supplies that constraint in a form usable here, and the reason is worth stating plainly, since it shapes what Chapter 5 can and cannot do. Independence between factors is the property one would like to enforce, but enforcing it directly requires knowing what the factors are. For synthetic benchmarks, the generative factors are known by construction. For natural images they are not, and no enumeration of them exists to supervise against. A constraint that is to work on real data must therefore act on some measurable

property of the representation that does not presuppose the identity of the factors it separates.

Disentangled representation learning aims to separate distinct factors of variation (e.g., pose, shape, texture) into independent latent dimensions. Within this area, probabilistic models such as Greff et al. (2020); Burgess et al. (2019) can achieve a degree of disentanglement due to their VAE backbone. Other works, such as Anciukevicius et al. (2020), pursued explicit disentanglement of position and depth, also within a probabilistic framework. Mansouri et al. (2023), instead, exploited weak supervision from sparse perturbations and causal representation learning to disentangle object properties. In a non-probabilistic setting, Singh et al. (2022) learned disentangled representations in an implicit manner, while Biza et al. (2023) introduced invariance to changes in position, scale, and rotation using slot-centric reference frames, thereby enabling the explicit disentanglement of these three factors.

Differentiation from Our Work: Our work Tang and Qi (2026) introduces the novel concept of "Factor Tokens" within a standard ViT. Unlike VAEs, which sacrifice generation quality for disentanglement (via the ELBO objective), our Minimum Volume Constraint (MVC) allows us to enforce geometric orthogonality in the latent space without degrading the discriminative power of the features, serving as a bridge to controllable generative models.

Chapter 3

Foundations in Remote Sensing: Solving Scale and Consistency Challenges

This chapter presents the foundational work of this dissertation, addressing fundamental challenges in self-supervised representation learning for remote sensing imagery. Remote sensing presents unique characteristics—particularly extreme scale variations and the need for both global semantic understanding and local spatial precision—that make it an ideal testbed for advancing representation learning methods. We begin by characterizing the remote sensing domain and its challenges, then present two contributions: global-local balanced contrastive learning for dense prediction tasks, and Cross-Scale MAE for multi-scale masked image modeling. These works establish core principles that subsequently generalize to natural images and inform later stages of this research.

3.1 The Remote Sensing Challenge

Remote sensing imagery, captured by satellites and aerial platforms, provides critical data for applications ranging from environmental monitoring and urban planning to disaster response and agricultural analysis. However, the unique characteristics of remote sensing data present substantial challenges for representation learning methods designed for natural images. The most prominent challenge is the extreme variation in ground sampling distance (GSD)—the physical distance on the ground represented by each pixel. Different imaging platforms operate at vastly different altitudes and sensor configurations, resulting in GSD ranging from sub-meter resolution (0.3m for very high resolution commercial satellites) to hundreds of meters for lower resolution sensors (250m-1km for MODIS). This creates a fundamental problem: objects of the same semantic class appear at dramatically different pixel scales depending on the GSD. A residential building in 0.3m GSD imagery might occupy 200×200 pixels with clearly visible roof details, while the same building type in 5m GSD imagery might occupy only 10×10 pixels, appearing as a small blob with no discernible internal structure. Large-scale datasets like FMOW Christie et al. (2018) contain imagery with GSD ranging from 0.3m to 15m—a 50× variation in effective resolution—making it challenging to train models that generalize across this scale variation. Beyond GSD variation, remote sensing scenes inherently contain multi-scale objects: an urban area at 0.5m GSD might simultaneously include individual vehicles (10 pixels), buildings (100 pixels), city blocks (1000 pixels), and entire neighborhoods. Effective understanding requires processing information at all these scales simultaneously, with spatial context playing a crucial role—a swimming pool adjacent to a large building suggests a hotel, while the same feature in a residential area indicates a private residence. Traditional multi-scale approaches like feature pyramid networks, designed for natural images with modest scale variations (2-4×), struggle with remote sensing's 10-50× variations.

Many critical remote sensing applications require dense, pixel-level predictions rather than image-level labels—semantic segmentation identifies land cover types at pixel resolution, building footprint extraction delineates individual structures, and change detection identifies precisely which pixels changed between time points. These tasks demand representations that preserve fine-grained spatial information and local consistency, not just global semantic understanding. This requirement poses challenges for self-supervised learning: contrastive methods like MoCo He et al. (2020a) and SimCLR Chen et al. (2020a) excel at image-level representations but may not preserve spatial precision needed for dense prediction, with aggressive cropping augmentations disrupting local spatial relationships. Masked image modeling approaches like MAE He et al. (2022b) naturally encourage learning spatial details through reconstruction, but standard MAE operates at a single scale and does not explicitly handle multi-scale variations or enforce learning relationships between different scales. Remote sensing imagery also differs from natural images in viewing geometry (predominantly nadir or near-nadir), spectral information (often extending beyond RGB to include infrared bands), and content distribution (background often dominates with objects of interest occupying small regions, highly imbalanced class distributions, and geographic/seasonal variations creating distributional shifts). These domain characteristics suggest that representation learning methods optimized for natural images require substantial adaptation. Understanding these challenges motivated our Phase 1 work: developing specialized approaches for remote sensing that establish principles subsequently generalizing to other domains.

3.2 Global-Local Balanced Contrastive Learning

This section presents our first contribution: a framework that explicitly balances global semantic understanding with local spatial consistency for improved dense prediction in aerial imagery Tang et al. (2023b). We begin by motivating the need for local consistency in contrastive learning, describe the framework architecture and

training methodology, present experimental results demonstrating effectiveness for pixel-level segmentation, and conclude with insights that inform subsequent work.

3.2.1 Motivation: The Local Consistency Problem

Standard contrastive learning methods achieve strong performance on image classification by learning representations that cluster semantically similar images while separating dissimilar ones. However, this global focus becomes problematic for dense prediction tasks requiring pixel-level accuracy. Semantic segmentation of aerial imagery requires representations that preserve spatial relationships and maintain local consistency—adjacent pixels with similar visual properties should have similar representations, even if they belong to different global semantic categories. The contrastive objective provides no explicit incentive to preserve spatial structure within the representation. Two images with similar global semantics (both urban scenes) might have very different spatial layouts (dense downtown vs. suburban sprawl), and standard contrastive learning correctly treats these as positive pairs for the global objective, providing no signal for learning spatially consistent representations needed for pixel-wise tasks. Data augmentation strategies exacerbate the problem: random cropping can disrupt spatial context by placing the same region in different relative positions across views, while color jittering may remove local textural information useful for distinguishing similar-looking categories. Prior work on dense contrastive learning Wang et al. (2021) applied contrastive objectives at the pixel level, but this requires dense correspondences and substantially increases computational cost. We sought a simpler approach that maintains global contrastive learning's efficiency while explicitly encouraging local spatial consistency through complementary objectives operating on spatial feature maps.

3.2.2 Framework Design

Our framework combines a global contrastive objective operating on image-level representations with a local consistency objective operating on spatial feature maps. The architecture consists of a shared encoder backbone, global and local projection heads, and complementary loss functions that balance global semantics with local spatial structure.

Architecture Overview

Given an input image $\mathbf{X} \in \mathbb{R}^{H \times W \times 3}$, we generate two augmented views $\mathbf{X}_1$ and $\mathbf{X}_2$ through random transformations including cropping, flipping, color jittering, and Gaussian blur (following SimCLR Chen et al. (2020a)). Both views are processed by a shared encoder $f_{\text{enc}}(\cdot)$ (we use ResNet-50 He et al. (2016) as the backbone) producing spatial feature maps:

$$\mathbf{F}_i = f_{\text{enc}}(\mathbf{X}_i) \in \mathbb{R}^{h \times w \times d}$$

where h and w are the spatial dimensions of the feature map (typically $H/32$ and $W/32$ for ResNet-50) and d is the feature dimension (2048 for ResNet-50).

Global Contrastive Branch

Feature-level CL. The feature-level contrastive learning module, $\boldsymbol{E}$, as shown in Fig. 3.1, uses the encoder of DeepLab v3+ with ResNet50 as backbone. Upon feeding an input image patch p to $\boldsymbol{E}$, a high-level representation, $\boldsymbol{E}(p)$, can be obtained. These representations are then used to generate the style feature Zhong et al. (2022), including both the channel-wise mean, $\mu(\cdot)$, and the variance, $\sigma(\cdot)$, as following,

$$\mathbf{f}(p) = concat(\mu(\boldsymbol{E}(p)), \sigma(\boldsymbol{E}(p))) \tag{3.1}$$

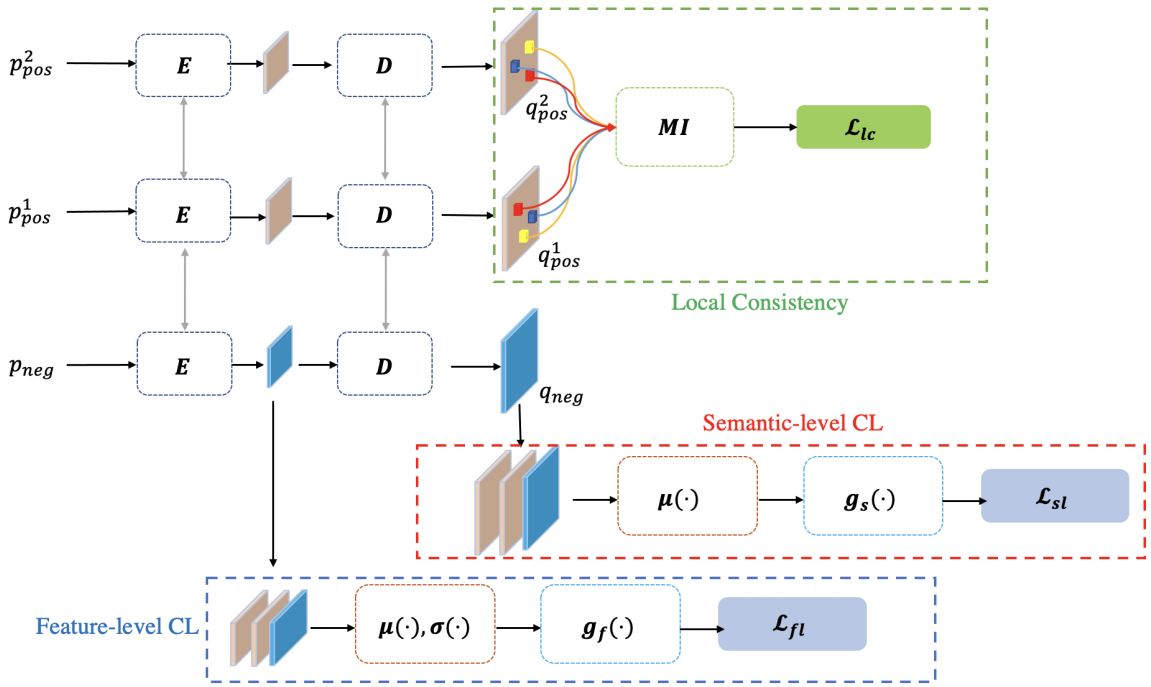


Figure 3.1: Illustration of the mCL-LC architecture.

Before computing the contrastive loss, a nonlinear projection head $\mathbf{g}_f(\cdot)$ is needed, which has been proven to be effective Chen et al. (2020c). So, after the encoder and projection head, we obtain the representation, $z = g_f(f(p))$. The feature-level contrastive loss is thus defined as:

$$\mathcal{L}_{fl} = \frac{1}{2N} \sum_{k=1}^{N} \left(\ell_{NTX}\left(\tilde{p}_k, \hat{p}_k\right) + \ell_{NTX}\left(\hat{p}_k, \tilde{p}_k\right) \right) \tag{3.2}$$

where $\hat{p}_k$ and $\tilde{p}_k$ are a positive patch pair generated from augmentations of the same patch p_k, and the NT-Xent contrastive loss function ℓ_{NTX} is the same as in SimCLR Chen et al. (2020b), which is defined as following:

$$\ell_{NTX}\left(\tilde{p}_k, \hat{p}_k\right) = -\log \frac{\exp\left(\text{sim}\left(\tilde{z}_k, \hat{z}_k\right)/\tau\right)}{\sum_{p \in \Lambda^-} \exp\left(\frac{\text{sim}\left(\tilde{z}_k, g_f(f(p))\right)}{\tau}\right)} \tag{3.3}$$

where Λ^- is the rest of the patch sets in the batch and $\tilde{z}_k = g_f\left(f\left(\tilde{p}_k\right)\right)$, $\hat{z}_k = g_f\left(f\left(\hat{p}_k\right)\right)$. Via minimizing the contrastive loss, it learns by forcing the representations from positive view pairs to be similar but those from negative pairs to be dissimilar.

Semantic-level CL. For segmentation tasks, a decoder structure is needed that learns by forcing the semantic feature to be similar in positive pair and dissimilar in negative pair. Here, we use the decoder of DeepLab v3+ as the semantic module, denoted as $\boldsymbol{D}$ (Fig. 3.1). Given the patch p_k, $\boldsymbol{D}$ receives the output of $\boldsymbol{E}$ and generates a semantic representation of the same size as that of p_k, which is denoted as $q_k = \boldsymbol{D}(\boldsymbol{E}(p_k))$. Integrating the channel information, we thus produce a pseudo semantic map s_k, and $s_k = g_s(\mu(q_k))$, where $\mu(\cdot)$ is the channel-wise average operator and $g_s(\cdot)$ is the projection head. Following the same procedure as in the feature-level contrastive loss, we apply the NT-Xent to calculate the semantic contrastive loss, which is defined as,

$$\mathcal{L}_{sl} = \frac{1}{2N} \sum_{k=1}^{N} \left(\ell_{NTX}\left(\tilde{s}_k, \hat{s}_k\right) + \ell_{NTX}\left(\hat{s}_k, \tilde{s}_k\right)\right) \tag{3.4}$$

$$\ell_{NTX}\left(\tilde{s}_k, \hat{s}_k\right) = -\log \frac{\exp\left(\operatorname{sim}\left(\tilde{s}_k, \hat{s}_k\right)/\tau\right)}{\sum_{p \in \Lambda_S^-} \exp\left(\frac{\operatorname{sim}(\tilde{s}_k, g_s(\mu(\boldsymbol{D}(\boldsymbol{E}(p)))))}{\tau}\right)} \tag{3.5}$$

where N denotes the number of patch pairs from a mini-batch of N samples, Λ_S^- is a set of pseudo maps corresponding to all patches except for the positive pair, and $g_s(\cdot)$ is a projection head similar to $g_f(\cdot)$.

Local Consistency Branch

We have constructed a multi-level contrastive learning framework. However, the complex acquisition conditions of remote sensing imagery demand more robust representation schemes that reveal the rich details hidden beneath the large footprint. In other words, a module that can understand the local semantic details is needed.

In Caron et al. (2020), the contrastive loss in the local region is used to improve the model's performance in learning representation of natural images, which forces similarity of local region of interests (ROIs) in positive pair, but dissimilarity in negative pair. However, this strategy is not adequate for aerial imagery because of its unique characteristics, i.e., spatial auto-correlation, that can cause two ROIs in different images to be similar. This is illustrated in Fig. 3.2, where two different images form a negative pair, but the three ROI pairs, matched by their geo-location, can be either similar (e.g., the red and yellow ROI pairs) or dissimilar (e.g., the blue ROI pair). However, the contrastive loss would have forced representations of similar patches to be dissimilar, which is undesirable. Hence we introduce the local consistency loss, to ensure the network only preserves the consistency of the local semantic information in the matching position of positive pairs.



Figure 3.2: ROIs in negative pair. The red, yellow, blue boxes are the matching ROIs in a negative image pair, but the matching ROIs bounded by the red and yellow boxes are actually similar.

Specifically, to obtain the matching ROIs in the positive pair, we first randomly select an ROI from $\tilde{p}$, then determine the location of the same size matching ROI in $\hat{p}$ according to the position of the ROI in $\tilde{p}$ to ensure the center of the two matching local regions point to the same position in the original image. By repeating these steps, we can select multiple different ROIs. We pass the locations of these ROIs to the last layer of $\boldsymbol{D}$ and select the corresponding ROIs in the pseudo-semantic map. Although

these matching local regions are from different augmentations, they should share the same content. We thus implement the consistency loss between the matching positive ROI pair, $\tilde{r}_j$ and $\hat{r}_j$, by calculating the mutual information and maximizing it.

Mutual information (MI) has been widely used for problems like multi-modality registration Zitová and Flusser (2003); Woo et al. (2015). It is a Shannon-entropy based measurement of mutual independence between two random variables, e.g., $\tilde{r}_j$ and $\hat{r}_j$. The mutual information $\mathcal{I}(\tilde{r}_j; \hat{r}_j)$ measures how much uncertainty of one variable ($\tilde{r}_j$ or $\hat{r}_j$) is reduced given the other variable ($\hat{r}_j$ or $\tilde{r}_j$). Based on the mutual information, the local consistency loss is defined as follows:

$$\mathcal{L}_{lc} = -\frac{1}{N_R} \sum_{j=1}^{N_R} \left(\mathcal{I}(\tilde{r}_j; \hat{r}_j) \right) \tag{3.6}$$

with

$$\mathcal{I}(\tilde{r}_j; \hat{r}_j) = H(\tilde{r}_j) - H(\tilde{r}_j \mid \hat{r}_j) \tag{3.7}$$

$$= \int_{\tilde{r}_j \times \hat{r}_j} \log \frac{\mathbb{P}_{\tilde{r}_j \hat{r}_j}}{\mathbb{P}_{\tilde{r}_j} \otimes \mathbb{P}_{\hat{r}_j}} d\mathbb{P}_{\tilde{r}_j \hat{r}_j}, \tag{3.8}$$

where H indicates the Shannon entropy and $H(\tilde{r}_j \mid \hat{r}_j)$ is the conditional entropy of $\tilde{r}_j$ given $\hat{r}_j$. N_R is the number of ROIs in one patch, and $\tilde{r}_j$ is the corresponding ROI in the pseudo semantic map. The exact value of the mutual information is difficult to calculate, so we resort to an estimate based on the MINE algorithm Belghazi et al. (2018), which is implemented by back propagation in a two-layer fully connected network.

Balanced Objective

To summarize, the final loss function of mCL-LC consists of three parts: 1) feature-level contrastive loss, $\mathcal{L}_{fl}$ (Eq. 3.2); 2) semantic-level contrastive loss, $\mathcal{L}_{sl}$ (Eq. 3.4); and 3) local consistency loss, $\mathcal{L}_{lc}$ (Eq. 3.6). The final loss function is defined as follows:

Table 3.1: COMPARISON WITH STATE-OF-THE-ART CONTRASTIVE LEARNING FRAMEWORKS IN TERMS OF PIXEL-BASED SEMANTIC SEGMENTATION ACCURACY. NOTE THAT SUPERVISED BASELINE REFERS TO DEEPLAB V3+

	Nice		Nantes Saint		Potsdam		Vaihungen	
	OA	Kappa	OA	Kappa	OA	Kappa	OA	Kappa
Supervised Baseline	0.6712	0.5410	0.6628	0.5312	0.7518	0.6812	0.7603	0.7037
SimCLR	0.6127	0.5013	0.6025	0.4413	0.7327	0.6472	0.7286	0.6304
MoCo v2	0.6277	0.5082	0.6201	0.4513	0.7371	0.6735	0.7309	0.6213
BYOL	0.6366	0.5230	0.6311	0.4850	0.7509	0.6715	0.7546	0.6407
SimSiam	0.6305	0.5022	0.6036	0.4735	0.7439	0.6903	0.7492	0.6333
GLCNet	0.6494	0.5302	0.6407	0.5291	0.7811	0.7179	0.7855	0.6807
mCL-LC	**0.6944**	**0.5379**	**0.6793**	**0.5343**	**0.8053**	**0.7301**	**0.8281**	**0.7377**
mCL-LC +	**0.7120**	**0.5520**	**0.7008**	**0.5689**	**0.8217**	**0.7440**	**0.8453**	**0.7506**

$$\mathcal{L} = \mathcal{L}_{fl} + \mathcal{L}_{sl} + \mathcal{L}_{lc}, \tag{3.9}$$

3.2.3 Experimental Validation

We evaluate the framework on multiple aerial imagery datasets and dense prediction tasks, focusing on semantic segmentation as the primary application requiring both global understanding and local spatial precision.

Comparison with State-of-the-Art

In this set of experiments, we evaluate the performance of mCL-LC from the perspectives of OA, Kappa, and F1-score, in comparison with five state-of-the-art contrastive learning frameworks. We also show the effectiveness of the pseudo-cloud noise generation augmentation method. The results are shown in Table 3.1, where "mCL-LC" means no pseudo cloud in the data augmentation, and "mCL-LC+" is with the pseudo cloud augmentation. From these results, we observe that mCL-LC outperforms all other contrastive learning frameworks in terms of both OA and

Table 3.2: ZERO-SHOT GENERALIZATION COMPARISON IN SITES TRANSFER (OA)

OA	SimCLR	MoCo v2	GLCNet	BYOL	SimSiam	mCL-LC
Nice → Nantes Saint	0.4402	0.4684	0.4744	0.4325	0.3926	**0.6108**
Nantes Saint → Nice	0.0816	0.1762	0.3018	0.2004	0.2131	**0.4719**

Kappa. The addition of pseudo-cloud augmentation further improves performance by roughly 2%.

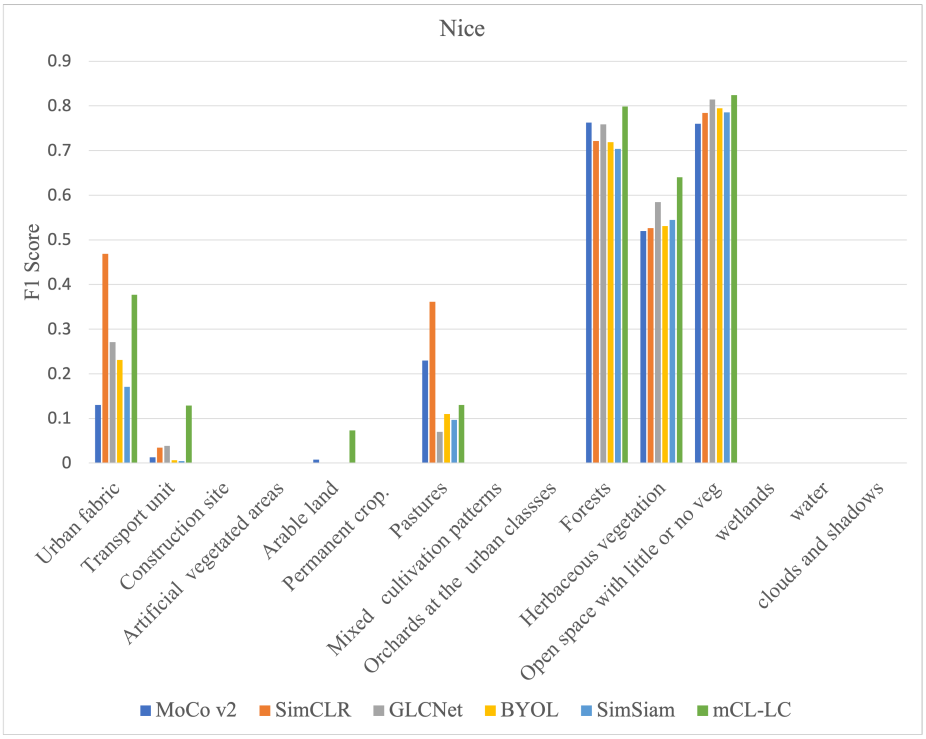
In addition to the OA and Kappa metrics, we also compute the F1-score for each category. These results are shown in Fig. 3.3. From the F1-score, we again observe that mCL-LC achieves the best performance across most categories.

Generalization Analysis

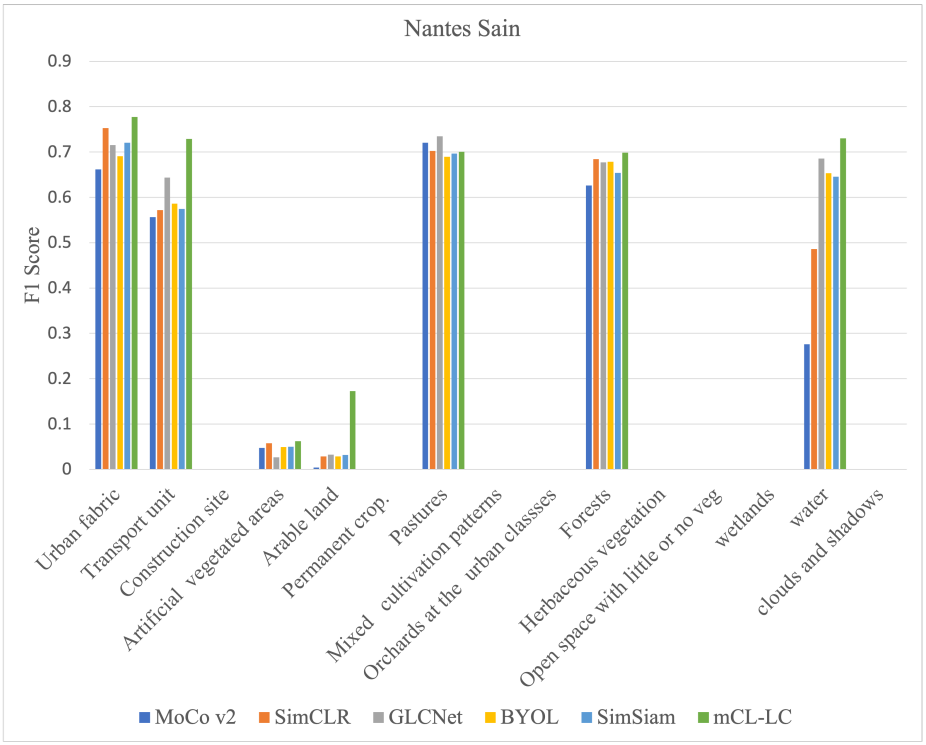
In this set of experiments, we evaluate the generalization capacity of mCL-LC using zero-shot domain testing. Specifically, training and fine-tuning are performed in one city, and testing is conducted in the other. The OA and Kappa for the different contrastive learning frameworks are shown in Table 3.2, where the city to the left of the arrow indicates the training and fine-tuning city, and that to the right is the testing city. From this table, we can observe that mCL-LC outperforms all other contrastive learning frameworks by more than 15%.

3.2.4 Discussion

The results in Table 3.1 support a narrow claim: when an image-level contrastive objective is balanced against a spatial-level consistency objective, dense prediction improves over either objective alone. mCL-LC exceeds all five contrastive baselines on both overall accuracy and Kappa across the four sites, and pretraining on aerial imagery outperforms ImageNet pretraining on every one of them, which is the evidence for preferring domain-specific pretraining here.



(a) Nice



(b) Nantes Saint

Figure 3.3: Class-wise F1-score on Nice and Nantes Saint.

The framework's scope is equally clear. It operates at a fixed spatial resolution, and on datasets that mix ground sampling distances, its accuracy falls substantially: representations learned at one scale do not transfer to another on their own. Scale is the form of heterogeneity mCL-LC does not address, and it is where the next section begins.

3.3 Cross-Scale MAE: Multi-Scale Masked Image Modeling

This section presents our second Phase 1 contribution: Cross-Scale MAE, a masked image modeling framework explicitly designed for extreme multi-scale scenarios in remote sensing Tang et al. (2023a). We begin by motivating the need for multi-scale MAE, describe the cross-scale architecture and masking strategies, present extensive experimental results across remote sensing benchmarks, and discuss fine-tuning strategies Georgiou et al. (2024).

3.3.1 Motivation: Multi-Scale Challenges in Masked Image Modeling

Remote sensing imagery exhibits extreme scale variations due to differences in ground sampling distance and the inherent presence of multi-scale objects within scenes. While global-local balanced contrastive learning addressed spatial consistency, it did not explicitly tackle the multi-scale challenge. Standard MAE He et al. (2022b), despite strong performance on natural images, faces limitations in multi-scale remote sensing scenarios. First, MAE operates at a fixed patch size (typically 16×16 pixels) throughout the encoder, corresponding to a fixed physical scale—objects appearing at different pixel scales due to GSD variation are processed with the same patch size, potentially missing scale-appropriate features. Second, there is no explicit mechanism ensuring scale-invariant representations; the model might learn separate

representations for "building at high resolution" versus "building at low resolution" rather than a unified scale-invariant concept. Third, the uniform random masking strategy treats all spatial regions equally, regardless of structure scale, while multi-scale imagery might benefit from different masking granularities for different regions. Fourth, MAE's encoder processes patches at a single resolution, lacking explicit mechanisms for processing the same spatial region at multiple scales simultaneously to capture both fine details and broad context. Recent work provided relevant context: SatMAE Cong et al. (2022) incorporated temporal and multi-spectral information but processed each image at single spatial resolution, while Scale-MAE Reed et al. (2022) introduced multi-scale processing through patches at different scales from the same spatial location. We build on these insights while introducing novel cross-scale masking strategies that enforce learning relationships between scales, explicit cross-scale reconstruction objectives encouraging scale-invariant features, scale-aware positional encodings, and extensive validation across diverse tasks demonstrating that these principles generalize beyond remote sensing.

3.3.2 Cross-Scale MAE Architecture

Our framework extends MAE with explicit multi-scale processing throughout the encoder, decoder, and masking strategy, as illustrated in Fig. 3.4. Its objective is to learn consistent representations from multi-scale images captured at the same site by harnessing the strengths of both discriminative and generative learning approaches. Cross-Scale MAE incorporates two novel loss functions at the encoder and decoder stages to achieve this objective. Firstly, the representations obtained from images of different scales are enforced to be consistent. This ensures that the learned representations capture relevant information *across* scales. Secondly, a decoder is employed to reconstruct the corresponding input, promoting the learned representation to be more representative. Additionally, the decoder uses cross-prediction, in which embeddings generated from an image with a lower GSD are

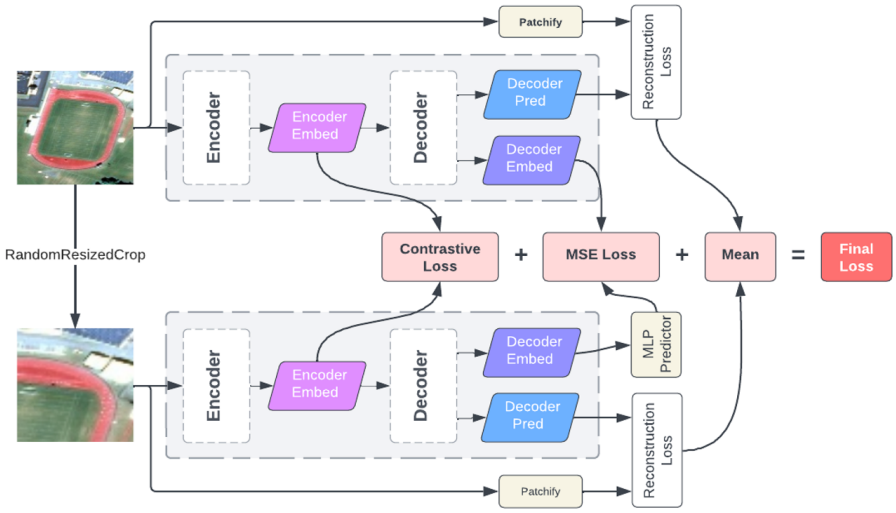
used to predict embeddings from an image with a higher GSD. This cross-prediction mechanism helps capture and leverage the multi-scale information present in the images.

Basic Setting

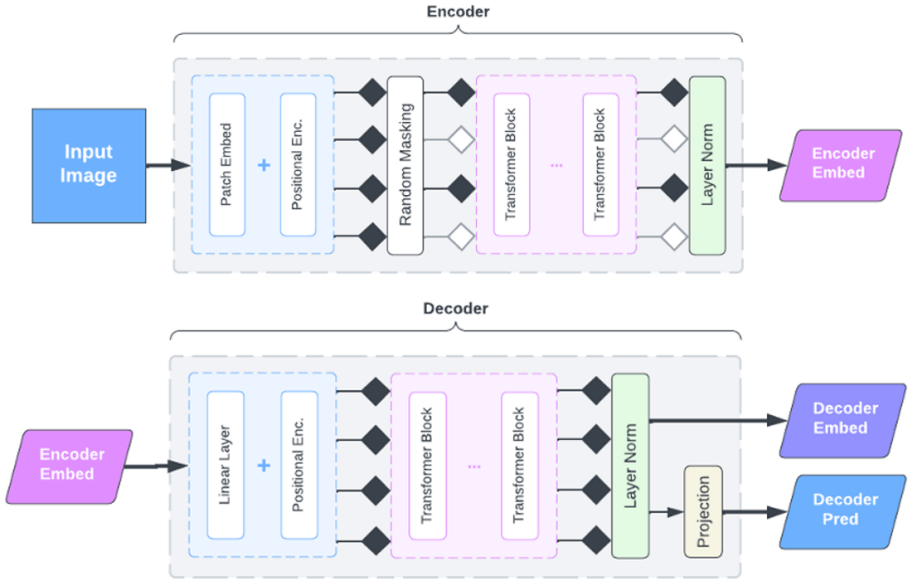
Let $p \in \mathbb{R}^{H \times W \times C}$ represent an input image of height H and width W with C channels. First, multi-scale augmentation (Sec. 3.3.2) is applied to p to generate two images, p_h of relatively higher GSD and p_l of relatively lower GSD. For each of the augmented images, a sequence of $|S|$ independent patches of height and width of n pixels are created, where each of the patches, $s \in S$, has dimension $s \in \mathbb{R}^{n^2 C}$. A fraction, m, of the patches is removed, while the remaining patches are passed through a projection function (e.g., a linear layer) to project the patches s into D dimensions, $f_{emd} : \mathbb{R}^{n^2 C} \rightarrow \mathbb{R}^D$, to obtain embedded patches $S_{emd} = f_{emd}(S)$. A standard positional encoding vector is added to the embedded patches and fed into the encoder. After the encoder, the removed m patches are placed back into their original location in the sequence of patches where a learned mask token represents the masked patches that were not encoded. Another positional encoding vector is added to all patches, and a sequence of transformer blocks decodes these patches to form the original input image; these steps are illustrated in Fig. 3.4.

Multi-Scale Augmentation

A specialized augmentation method is applied to the input image before patchifying it in the Cross-Scale MAE framework. This augmentation process generates an additional image of lower GSD for the same site, and the two images are then used as inputs through two independent instances of the MAE network. More specifically, given an input image p with a certain GSD, the framework randomly selects two scale ratios, r_h and r_l, from the range [0, 1], where $r_l < r_h$. The image p is then down-sampled using these scale ratios to generate two images, p_h and p_l. In our



(a) Full Pipeline



(b) Sub-Modules

Figure 3.4: The architecture of Cross-Scale MAE comprises an encoder and a decoder (a). The encoder (b-top) employs a vision transformer (ViT) backbone, specifically a ViT-base with 12 self-attention blocks. The decoder (b-bottom) uses a lightweight ViT backbone with 8 self-attention blocks. A single satellite image undergoes scale augmentation through random cropping and resizing. The contrastive loss is computed using encoder outputs for the two scale inputs. The cross-prediction loss (in MSE) is applied to the decoder output of the last self-attention block. The reconstruction loss compares the predicted masked patch with the actual input.

experiments, we set $r_h = 1$, so $p_h = p$. Both images, p_h and p_l, are subsequently resized to a fixed size of 128px^2, and the GSD values of p_h and p_l are denoted as g_h and g_l, respectively.

Note that this augmentation approach differs from traditional augmentation, where the rescaled images are traditionally only used as separate samples. Instead, the multi-scale inputs are generated dynamically at runtime and used as inputs through multiple network branches, where the branches have shared weights. The constraints enforced upon the learned representations of these network branches are detailed in the following sections.

Cross-Scale Information Consistency in the Encoder

Typically, the standard MAE trains a network to reconstruct an image by masking out a significant portion of its pixels. However, the standard MAE encoder does not explicitly address the issue of representation consistency for multi-scale input sources. In contrast, Cross-Scale MAE aims to learn meaningful and consistent scale-invariant representations. To achieve this, Cross-Scale MAE focuses on two fundamental properties in the learned representations: (1) consistency of information across different-scale images and (2) representativeness of the raw image. In the Cross-Scale MAE framework, a contrastive learning strategy is employed within the encoder to ensure consistency among representations derived from different-scale images.

The intuitive idea behind Cross-Scale MAE is that different-scale images captured at the same ground site contain highly relevant information despite potential visual differences. Cross-Scale MAE leverages this notion by maximizing the shared information between representations derived from different-scale images of the same site while minimizing the shared information between images from different sites. This approach aligns with the principles of contrastive learning. In contrastive learning, the InfoNCE loss is commonly used to estimate the lower bound of mutual information between data samples. By minimizing the InfoNCE loss, we effectively maximize the

mutual information between representations from different scales Tian et al. (2020); Oord et al. (2018).

Specifically, the encoder E, uses the ViT-base Dosovitskiy et al. (2020) as the backbone. Upon feeding an input image p to encoder E, a high-level representation, $E(p)$, can be obtained. Before computing the contrastive loss, a nonlinear projection head $g_f(\cdot)$ is needed, which has been proven to be effective Chen et al. (2020c). So, after the encoder and projection head, we obtain the representation, $z = g_f\left(E\left(p\right)\right)$. The cross-scale consistency loss is thus defined as:

$$\mathcal{L}_{cc} = \frac{1}{2N} \sum_{k=1}^{N} \left(\ell_{info}\left(p_l^k, p_h^k\right) + \ell_{info}\left(p_h^k, p_l^k\right) \right) \tag{3.10}$$

where p_h^k and p_l^k are the different-scale images from the same input image p_k, N denotes the number of samples in a mini-batch, and the InfoNCE contrastive loss function ℓ_{info} is the same as in Chen et al. (2020b), which is defined as follows:

$$\ell_{info}\left(p_l^k, p_h^k\right) = -\log \frac{\exp\left(\text{sim}\left(z_l^k, z_h^k\right)/\tau\right)}{\sum_{p\in\Lambda} \exp\left(\frac{\text{sim}\left(z_l^k, g_f(E(p))\right)}{\tau}\right)} \tag{3.11}$$

where $z_l^k = g_f\left(E\left(p_l^k\right)\right)$, $z_h^k = g_f\left(E\left(p_h^k\right)\right)$ and Λ is the sample batch. The sim operator is the cosine distance.

Cross-Scale Prediction in the Decoder

The InfoNCE loss in the encoder provides consistency across multi-scale inputs, but cannot guarantee that the learned representation adequately captures the input. In other words, the representation learned by the network from multi-scale images should not only be consistent across scales but also representative of semantic information. According to Caron et al. (2021), the self-supervised ViT automatically learns class-specific features leading to unsupervised object segmentation; specifically, the attention block's last layer includes abundant semantic information. Motivated

by Caron et al. (2021), the decoder of the Cross-Scale MAE has two tasks: (1) reconstruction of the multi-scale input as close as possible, (2) cross-scale prediction between attentions from different scales, as shown in the decoder part of Fig. 3.4 and detailed below.

Similar to the standard MAE, Cross-Scale MAE also adopts a light decoder. Decoding follows the standard MAE decoder, where, following the encoder, the removed m patches are placed back into their original location in the sequence of patches, where a learned mask token represents the masked patches that were not encoded, and a positional encoding is added. Then, a series of transformer layers decode all patches. Given the two representations $f_{e,h}$ and $f_{e,l}$, where $f_{e,h} = E(p_h)$, $f_{e,l} = E(p_l)$, after the decoder attention block, we will obtain two new representations, $f_{d,h}$ and $f_{d,l}$. Then, the cross-scale prediction is applied between them to get the cross-scale prediction loss, defined as:

$$\mathcal{L}_{cp} = \frac{1}{N} \sum_{k=1}^{N} \left(\ell_{pred} \left(f_{d,l}^{k}, f_{d,h}^{k} \right) \right) \tag{3.12}$$

where N is the batch size and the prediction loss ℓ_{pred} is defined as:

$$\ell_{pred} \left(f_{d,l}^{k}, f_{d,h}^{k} \right) = ||f_{d,h}^{k} - g_p(f_{d,l}^{k})||_2^2 \tag{3.13}$$

It follows the Mean Squared Error (MSE) between the prediction and target representation, where the predictor, $g_p(\cdot)$, is a multi-layer perceptron (MLP). In addition to the cross-scale prediction loss, the decoder reconstructs images at different scales. The reconstruction loss is thus defined as:

$$\mathcal{L}_{re} = \frac{1}{N} \sum_{k=1}^{N} \left(||p_l - \tilde{p}_l||_2^2 + ||p_h - \tilde{p}_h||_2^2 \right) \tag{3.14}$$

where $\tilde{p}_l = D(E(p_l))$ and $\tilde{p}_h = D(E(p_h))$. Through the cross-scale prediction loss and the reconstruction loss, the consistency and effectiveness of the representation's semantic information are better preserved.

Table 3.3: The information of datasets used in Sec. 3.3.3

Dataset	GSD(m)	Number of Images	Number of Categories
RESISC45	0.2-30	31500	45
WHU-19	0.5	1050	19
UC Merced	0.3	2100	21
EuroSAT	10	27000	10

The total loss of the Cross-Scale MAE is the sum of cross-scale consistency loss at the encoder ($\mathcal{L}_{cc}$), cross-scale prediction loss at the decoder ($\mathcal{L}_{cp}$), and the reconstruction loss ($\mathcal{L}_{re}$), defined as:

$$\mathcal{L} = \mathcal{L}_{cc} + \mathcal{L}_{cp} + \mathcal{L}_{re} \tag{3.15}$$

3.3.3 Experimental Results

We evaluate Cross-Scale MAE on multiple remote sensing benchmarks spanning classification and semantic segmentation, together with reconstruction quality on natural images. Evaluating across tasks that stress different properties of the representation, linear separability in one case and dense spatial precision in the other, tests whether the gains are general or specific to a single objective.

Comparison with State-of-the-Art

Similar to Scale-MAE, we use both the K-nearest neighbors (KNN) classification accuracy and performance in downstream tasks as metrics to evaluate the quality of the representation learned by Cross-Scale MAE. Specifically, we pre-train the Cross-Scale MAE on the fMoW-RGB dataset and freeze the encoder as a representation generator; we then evaluate its effectiveness on different datasets using KNN classification.

KNN Performance. Similar to the original MAE configuration, the encoder of Cross-Scale MAE uses a ViT-Large model, and the decoder is a light ViT model

with the 8-layer attention blocks. The datasets we use include RESISC45 Cheng et al. (2017a), WHU-RS19 Xia et al. (2017), UC Merced Xia et al. (2017), EuroSAT Helber et al. (2019), as shown in Table 3.3. To evaluate the representation learning capacity of Cross-Scale MAE with different GSD images, we evaluate the network performance using images with different scale ratios, $\{12.5\%, 25\%, 50\%, 100\%\}$, to the raw image to simulate different GSD.

The results are presented in Fig. 3.5. This figure shows the performance of Cross-Scale MAE compared to both SatMAE and Scale-MAE for different scale ratios. From Fig. 3.5, we observe that, in general, the Cross-Scale MAE performs better than SatMAE or Scale-MAE in all different scale ratios on all datasets. And, for scale ratio in $\{25\%, 50\%, 100\%\}$, the Cross-Scale MAE can obtain more stable performance compared with SatMAE or Scale-MAE. When the scale ratio equals 12.5%, all three models perform relatively worse. This is because the scale ratio set during pre-training is from 0.2 to 0.8, and 0.125 is out of this range. Nonetheless, Cross-Scale MAE still presents overwhelmingly better performance than the other two. Table 3.4 shows the average accuracy at all scale ratios and compares Cross-Scale MAE with SatMAE and Scale-MAE. We can find that Cross-Scale MAE presents overwhelmingly better performance than SatMAE in all datasets and we observe similar trend as in Fig. 2. Cross-Scale MAE generally performs better than Scale-MAE except in the UC Merced dataset. It may be because RESISC45 covers an extensive range of GSD, from $0.2m$ to $30m$, but the other three datasets have fixed GSD. This comparison effectively demonstrates the robustness of representations generated from Cross-Scale MAE, especially on multi-scale datasets.

Downstream Tasks. We further evaluate the performance of Cross-Scale MAE on different downstream tasks to assess the effectiveness of the learned representations in various practical scenarios, including classification and segmentation. We conduct the classification task on the fMoW-RGB dataset. The segmentation downstream tasks are performed on two datasets, Potsdam and Vaihungen. We fine-tune the

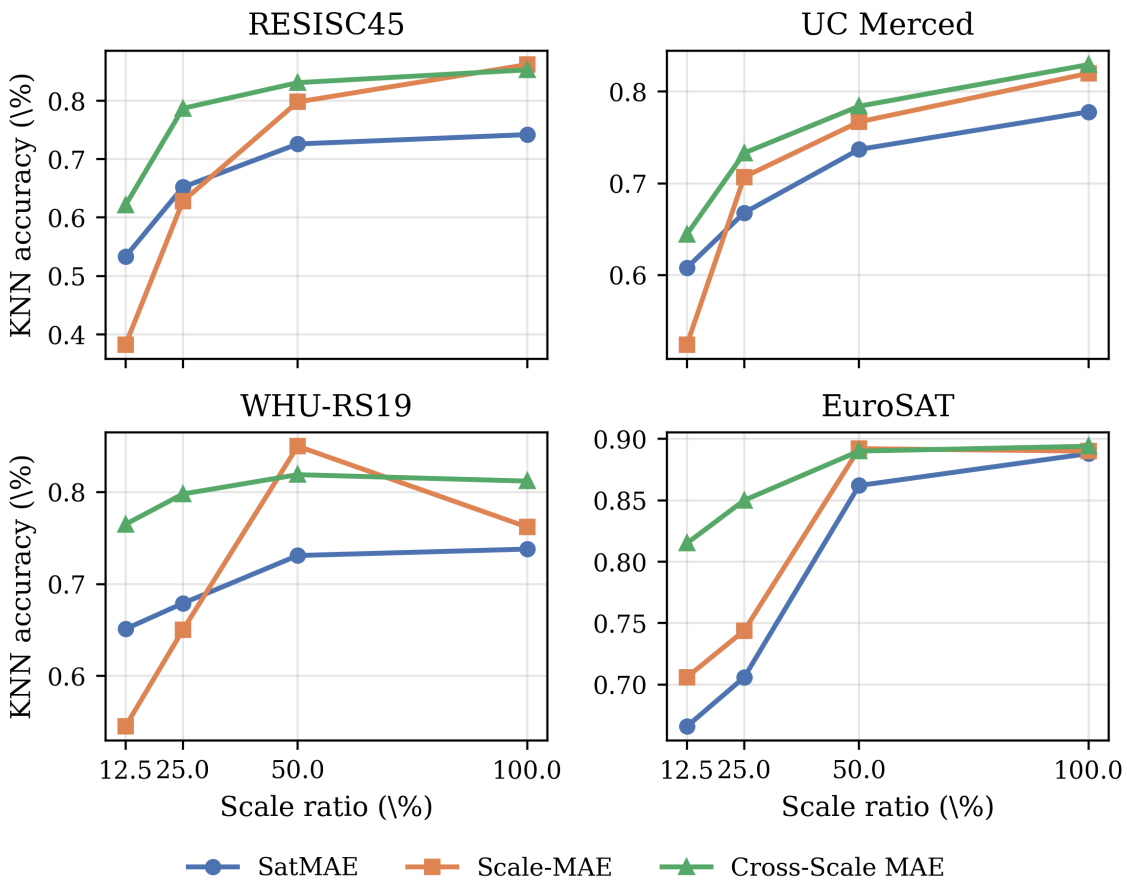


Figure 3.5: The KNN classification of Cross-Scale MAE for different datasets.

model for 50 epochs on all downstream tasks, using the same hyperparameter settings as in Scale-MAE Reed et al. (2022).

To test the capability of handling multi-scale inputs, we compare the performance of different methods on low-resolution images and on full-resolution images with varying scaling ratios. In Table 3.5, we report the classification performance regarding Top-1 and Top-5 accuracy. In Table 3.6, we report the segmentation performance regarding mIoU. Cross-Scale MAE outperforms state-of-the-art methods in both classification and semantic segmentation downstream tasks.

More ablation study results, please check our published paper Cross-Scale MAE Tang et al. (2023a).

Table 3.4: Average KNN accuracy with different scale ratios $(12.5\%, 25\%, 50\%, 100\%)$

	RESISC45	WHU-RS19	UC Merced	EuroSAT
SatMAE	66.3	69.9	69.7	81.9
Scale-MAE	70.0	79.5	75.0	86.7
Cross-Scale MAE	75.6	79.8	74.5	87.8

Table 3.5: Linear classification fine-tuning performance on fMoW-RGB

		Scale=50%		Scale=100%	
	Backbone	Top-1	Top-5	Top-1	Top-5
GASSL	ResNet50	0.494	0.540	0.683	0.715
SeCo	ResNet50	0.508	0.602	0.614	0.796
SatMAE	ViT-Base	0.551	0.776	0.651	0.892
SatMAE	ViT-Large	0.591	0.793	0.678	0.923
Scale-MAE	ViT-Base	0.623	0.850	0.673	0.930
Cross-Scale MAE	ViT-Base	**0.632**	**0.914**	**0.692**	**0.954**

Table 3.6: Semantic segmentation performance on Potsdam & Vaihingen (mIoU)

		Potsdam			Vaihingen		
	Backbone	25%	50%	100%	25%	50%	100%
Supervised Baseline	ResNet50	0.4518	0.5212	0.7411	0.4603	0.5237	0.7438
SimCLR	ResNet50	0.5327	0.5472	0.7233	0.5286	0.5904	0.7075
BYOL	ResNet50	0.5509	0.5715	0.7264	0.5618	0.5903	0.7298
SimSiam	ResNet50	0.5439	0.5903	0.7188	0.5522	0.6005	0.7205
SeCo	ResNet50	0.5571	0.5935	0.7298	0.5741	0.6213	0.7250
mCL-LC	ResNet50	0.5589	0.5973	0.7316	0.5795	0.6277	0.7262
SatMAE	ViT-Base	0.6325	0.6879	0.7355	0.6239	0.6980	0.7364
Scale-MAE	ViT-Base	0.6853	**0.7333**	0.7507	0.6910	0.7280	0.7512
Cross-Scale MAE	ViT-Base	**0.7082**	0.7225	**0.7617**	**0.7163**	**0.7354**	**0.7603**

Evaluation on Natural Images (CoCo Dataset)

To assess the generalization capabilities of Cross-Scale MAE, we evaluate by pre-training on the CoCo2017 dataset, focusing on natural images. This evaluation allows us to validate the effectiveness of our model beyond remote sensing images and examine its performance in a different domain. We pre-train the Cross-Scale MAE on the CoCo2017 dataset using the following parameter settings: ViT-Base as the backbone architecture, a base learning rate of 5×10^{-3}, a weight decay of 5×10^{-3}, and an input size of 128×128. The model is trained for 400 epochs.



Figure 3.6: Reconstruction performance on the CoCo Dataset

We visualize the reconstruction results on the CoCo dataset in Fig.3.6 and present the corresponding evaluation metrics in Fig.3.7. We compare the performance of Cross-Scale MAE with the baseline model, MAE He et al. (2022b). From Fig.3.7,

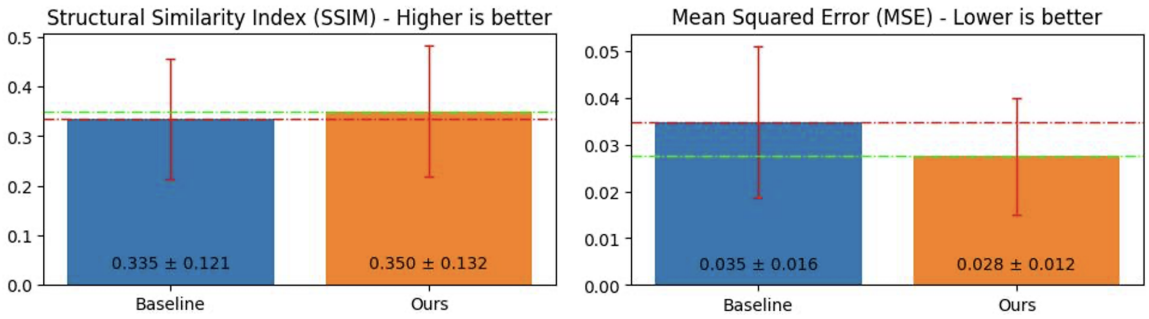


Figure 3.7: Multiscale reconstruction performance on the CoCo dataset

we observe that Cross-Scale MAE outperforms the baseline model in terms of both MSE and SSIM. Additionally, in Fig.3.7, we notice that the baseline model exhibits more artifacts in the reconstruction results at the locations of the masked patches. In contrast, Cross-Scale MAE provides a closer match to the actual distribution of pixel values expected at those locations.

Furthermore, we evaluate the zero-shot reconstruction performance of Cross-Scale MAE on the fMoW-RGB dataset using the model pre-trained on CoCo. In this evaluation, we freeze the model trained on CoCo and reconstruct images from the fMoW-RGB dataset. It is important to note that the model has not seen any images from the fMoW-RGB dataset during its training. The zero-shot reconstruction results are displayed in Fig. 3.8. These results indicate that the learned representations by Cross-Scale MAE generalize well, as the zero-shot reconstruction still produces meaningful outputs on the fMoW-RGB dataset.

The evaluation of natural images demonstrates the versatility of Cross-Scale MAE, showcasing its ability to capture meaningful representations and generalize effectively across different domains. These findings highlight the potential of our model to enhance image understanding and reconstruction tasks in various applications beyond remote sensing imagery.

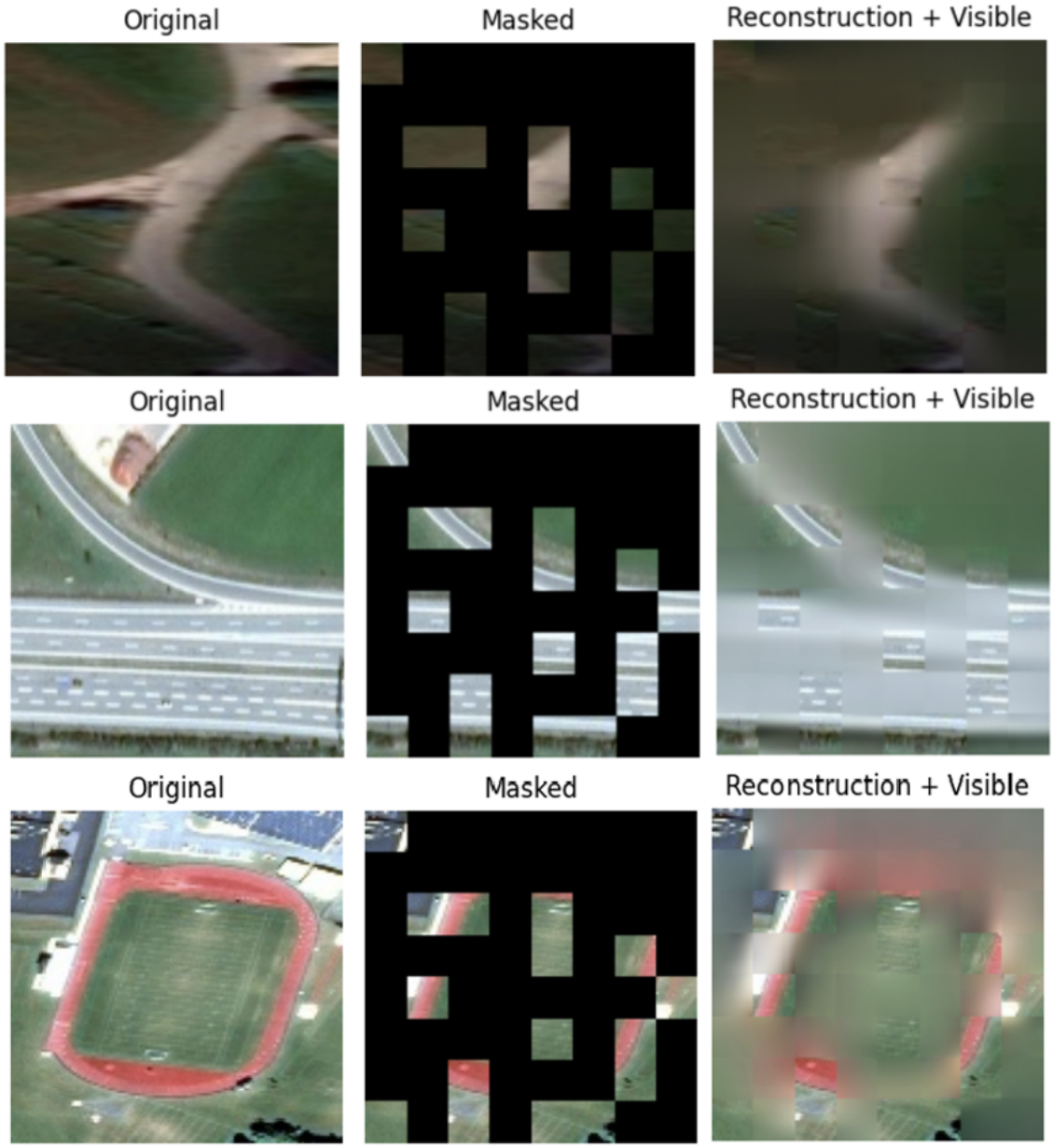


Figure 3.8: Zero-shot reconstruction in fMoW with Cross-Scale MAE pre-trained on CoCo

3.3.4 Fine-Tuning Validation: The SMART Dataset

A natural question following any self-supervised pretraining framework is how readily its representations adapt to a genuinely difficult downstream task with limited labeled data. A collaborative follow-up study Georgiou et al. (2024) (contributing author) investigated this question by fine-tuning Cross-Scale MAE for construction-phase classification on the IARPA-sponsored SMART dataset, which poses two challenges beyond those already discussed: imagery drawn from three distinct sensors (Landsat-8, Sentinel-2, and WorldView) at different native resolutions, and construction phases (site preparation, active construction, post-construction, and no activity) whose visual boundaries are inherently ambiguous.

Fine-tuning the pretrained Cross-Scale MAE encoder on high-resolution World-View imagery reached 68.5% overall accuracy, exceeding a fully supervised baseline trained on the same data (61.4%) despite requiring substantially less labeled data. Downsizing lower-resolution imagery to match the pretraining resolution consistently outperformed center-cropping, suggesting that preserving the full scene during resizing matters more than preserving native pixel density. A two-stage fine-tuning strategy that further adapted the model to co-registered multi-sensor pairs did not meaningfully improve over single-sensor fine-tuning, indicating a practical limit to how much additional adaptation stages contribute once the dominant challenge, resolution mismatch, has been addressed. Together, these results provide external validation that Cross-Scale MAE's pretrained representations transfer to a real downstream task with practical labeling constraints, beyond the benchmark evaluations above.

3.3.5 Discussion

Cross-Scale MAE improves scale robustness through explicit cross-scale patch extraction, masking, and reconstruction. It outperforms single-scale baselines on classification and semantic segmentation, and the ablation in Section 3.3.3 attributes most of that margin to the cross-scale masking strategy specifically: mixing aligned

masking across scales with scale-specific masking supplies a learning signal that within-scale masking alone does not.

Three limitations bound the result. Processing multiple scales costs roughly 3× single-scale MAE. Extrapolating well beyond the training scale range still degrades performance, so what the model acquires is scale robustness rather than scale invariance. And scale is not separated from content in the learned representation: a model that encoded "what object" independently of "at what scale" would generalize differently, and Cross-Scale MAE does not do that. The first two are costs; the third is a question the remaining chapters return to.

3.4 Summary

This chapter presented two contributions to self-supervised representation learning for remote sensing. Global-local balanced contrastive learning balances an image-level semantic objective against a spatial-level consistency objective, improving dense prediction where either alone falls short. Cross-Scale MAE addresses scale directly through multi-scale patch extraction, masking that enforces relationships between scales, and multi-scale reconstruction.

Both results point the same way. Neither the local detail that mCL-LC preserves nor the cross-scale consistency that Cross-Scale MAE maintains appeared on its own when the corresponding mechanism was removed; each had to be built into the objective. Domain characteristics were what made these mechanisms worth building, and self-supervised pretraining on aerial imagery outperformed transfer from natural images throughout.

That raises the question the next chapter takes up. Scale variation of 10 to 50× is particular to satellite imagery, but treating heterogeneous content non-uniformly is not. Natural images pose the same problem in a different form: not variation in the scale at which a patch was captured, but variation in which learning signal best serves it. Chapter 4 tests whether the principle survives that change of setting.

Chapter 4

Generalizing Consistency: Adaptive Loss Routing for Multi-Objective Masked Image Modeling

This chapter extends the consistency principle established in Phase 1 to a new setting. Chapter 3 showed that robust representations, whether balancing local and global structure or reconciling views captured at different scales, come from explicit, non-uniform treatment of heterogeneous content, not from processing applied uniformly everywhere. Phase 2 asks whether this same principle governs a different axis of heterogeneity: which learning signal a given patch should be trained with, when several complementary objectives are combined into one masked image modeling framework. This work, accepted at ECCV Georgiou et al. (2026), shows that repurposing Mixture-of-Experts routing as a mechanism for learned, per-patch loss weighting improves representation quality on ImageNet and ADE20K, extending the case for explicit, structure-aware mechanisms beyond the spatial and scale

heterogeneity of Phase 1, while revealing a limitation that carries forward into Phase 3: high performance does not guarantee an interpretable structure.

4.1 Generalizing the Consistency Principle

Both contributions in Phase 1 solved their respective problems through an explicit mechanism for consistency. Global-local balanced contrastive learning enforced consistency between an image's global semantic content and its local spatial detail, so that neither was learned at the expense of the other. Cross-Scale MAE enforced consistency between representations of the same site captured at different scales, so that a structure's representation under high and low ground sampling distance remained related rather than arbitrary. In both cases, achieving this consistency required treating different parts of the input, local versus global, high-GSD versus low-GSD, non-uniformly, through an explicit training objective, rather than hoping a single encoder would discover the relationship on its own.

This raised a natural question for Phase 2: is explicit, non-uniform treatment a principle specific to spatial and scale heterogeneity, or does the same lesson, that uniform treatment of heterogeneous content leaves performance on the table, extend to an entirely different axis of heterogeneity? Masked image modeling (MIM) offers such a setting. Following MAE's success He et al. (2022b), numerous variants emerged proposing different design choices: masking strategies (random He et al. (2022b), block-wise Xie et al. (2022), attention-based), reconstruction targets (pixels He et al. (2022b), discrete tokens Bao et al. (2021), teacher features Wei et al. (2022)), and teacher network selection (CLIP Radford et al. (2021), DINO Caron et al. (2021), supervised models). Nearly all frameworks that combine several of these learning signals do so with a single global scalar weight per objective, applied uniformly across every patch in every image, regardless of what that patch actually contains. Where Phase 1's heterogeneity was spatial and physical (scale), this heterogeneity is informational: some patches are semantically rich, an object's most distinctive parts,

while others are largely texture or background, and a uniform weighting scheme cannot express that these different patches are best served by different learning signals.

This observation motivated collaborative work testing whether Phase 1's principle extends to this new axis. As a contributing author (second author, equal contribution), the aspects presented here focus on the routing and loss-weighting design and its connection to this dissertation's broader argument: that consistent, high-quality representations come from making heterogeneity explicit in the training objective, not from treating every patch alike and hoping the network compensates. The goal was twofold: test whether the Phase 1 principle generalizes beyond scale specifically, and identify a concrete mechanism for expressing patch-level heterogeneity directly within the loss function.

4.2 Patch-Level Loss Routing for Multi-Objective Masked Image Modeling

This section presents the resulting framework, ExPLoRe (Expert Patch-Level Loss Routing) Georgiou et al. (2026). We focus on the aspects most relevant to the dissertation narrative: why global loss weighting is insufficient, how expert routing is repurposed to solve the problem, and what the experiments reveal about when and why this works.

4.2.1 The Patch-Level Heterogeneity Problem

Multi-objective MIM combines complementary learning signals to train a single student encoder: token-level distillation from a frozen teacher (here CLIP Radford et al. (2021)), global alignment between student and teacher `[CLS]` representations, and pixel-level reconstruction of masked patches. Each objective captures different information: token distillation transfers the teacher's semantic structure, CLS

alignment enforces global consistency, and pixel reconstruction preserves low-level detail. Prior work established that combining these objectives outperforms any single one in isolation, but left open the question of how they should be combined.

Existing multi-task approaches, including uncertainty weighting and gradient-normalization schemes such as GradNorm Chen et al. (2018), all operate at the global level: each objective receives one scalar weight applied uniformly across every spatial location. This ignores the spatial structure of the problem. Semantically rich patches are better served by teacher distillation, which carries high-level semantic information, while texture-heavy or background patches benefit more from pixel reconstruction, which directly supervises low-level appearance. A global weight cannot express this distinction, and gradient-based balancing can fail outright when objectives have spatially heterogeneous importance: in our experiments, GradNorm could not match its own static baseline on a token-distillation-plus-CLS-alignment objective (72.8% vs. 74.2% k-NN) and collapsed entirely once pixel reconstruction was added (32.5% k-NN). To our knowledge, no prior MIM work addresses loss weighting at the level of individual patches.

4.2.2 Repurposing Expert Routing as Per-Patch Loss Coefficients

The key idea is to repurpose Soft Mixture-of-Experts (Soft-MoE) routing Puigcerver et al. (2024), originally designed for conditional computation and capacity scaling, as a mechanism for learned, per-patch loss weighting. Soft-MoE replaces selected MLP layers with E expert networks and computes two distinct sets of weights from a shared set of routing logits: dispatch weights $\mathbf{D}$, normalized over patches for each expert ($\sum_n D_{n,e} = 1$), which determine how patches are aggregated into each expert's input, and combine weights $\mathbf{C}$, normalized over experts for each patch ($\sum_e C_{n,e} = 1$), which determine how expert outputs are recombined in the forward pass. Conventional Soft-MoE uses only the combine weights.

The central innovation is to repurpose the dispatch weights $\mathbf{D}$, not the combine weights, as per-patch loss coefficients. If Expert 0 is assigned to weight the token-distillation loss, for example, its contribution for image b becomes a dispatch-weighted average over visible patches $\mathcal{V}$:

$$\mathcal{L}_{\text{token}}^{\text{MoE}} = \frac{1}{B}\sum_{b=1}^{B} \frac{\sum_{i\in\mathcal{V}} D_{b,i,0} \cdot \ell_{\text{huber}}\big(\hat{\mathbf{t}}_{b,i}, \text{LN}(\mathbf{t}_{b,i})\big)}{\sum_{i\in\mathcal{V}} D_{b,i,0}} \tag{4.1}$$

where the per-image normalization in the denominator prevents the router from trivially minimizing the loss by shrinking weights uniformly across the batch. Analogous terms weight the CLS-alignment and pixel-reconstruction objectives when additional experts are used.

Using dispatch weights rather than combine weights is not an arbitrary choice. Combine weights are normalized *per patch* across experts, so a router could trivially minimize the weighted loss on a difficult patch by driving that patch's weight toward zero for every expert; used this way, loss weighting collapses representation quality to 2.1% k-NN, a 73-point drop, confirming this is a genuine degeneracy and not a minor inefficiency. Dispatch weights, normalized *per expert* across patches, cannot collapse this way: each expert must distribute its attention somewhere across the batch, so no single patch can be silently excluded from every loss term.

Because dispatch weights are differentiable, gradients from each loss objective flow back through $\mathbf{D}$ to the router parameters, a *loss-coupling* mechanism that lets the router learn which patches should receive which emphasis directly from the training signal, with no explicit supervision of what "semantically rich" or "texture-heavy" means. Because dispatch weights alone do not guarantee balanced expert utilization, a per-expert entropy regularizer discourages one expert from dominating all patches while another receives none; its strength trades off against loss-coupling and must be tuned per expert count, discussed next.

4.2.3 Mechanism Validation and Expert Scaling

We evaluate on ImageNet-1K with a ViT-Base student and a frozen CLIP ViT-B/16 teacher, replacing the MLP in alternating encoder blocks with Soft-MoE layers, under two objective combinations: Token+CLS and Token+Pixel. Table 4.1 isolates the contribution of each design choice on the Token+CLS setting. Two-expert MoE without loss-coupling underperforms the non-MoE baseline (74.1% vs. 75.7% k-NN); additional expert capacity alone is actively harmful without a mechanism to make use of it. A detach ablation, which blocks loss gradients from reaching the router while leaving dispatch weights in the forward pass, degrades performance by 1.6 points relative to the full mechanism, confirming that it is the router's access to gradient signal from the loss, and not merely the presence of experts, that drives the improvement. Scaling to 64 experts with the full mechanism reaches 76.2% k-NN and 80.6% linear-probe accuracy, a +4.4 point linear-probe gain over the non-MoE baseline.

Table 4.1: Mechanism validation on the Token+CLS configuration (ImageNet-1K, k-NN@20). Repurposing dispatch weights as loss coefficients, with gradients coupled to the router, is what turns additional expert capacity into a benefit rather than a liability.

Configuration	**k-NN@20 (%)**
No MoE (baseline)	75.7
2-expert MoE, combine-weight loss coefficients	2.1
2-expert MoE, no loss-coupling (unweighted)	74.1
2-expert MoE, dispatch-weight loss-coupling (ours)	75.4
64-expert MoE, dispatch-weight loss-coupling (ours)	**76.2**

The entropy-regularization weight λ governs a sharp transition rather than a gradual trade-off: reducing λ from 0.1 (stable, 75.3% k-NN) to 0.01 (collapsed, 66.9% k-NN) causes the dispatch-weight coefficient of variation across patches to jump by a factor of 32, indicating a degenerate shortcut rather than meaningful specialization.

The optimal value is expert-count dependent: $\lambda = 0.5$ is best for 2 experts but degrades 64-expert performance, since larger expert counts already provide enough inherent diversity that heavy regularization becomes counterproductive. A qualitative signature of successful specialization is that, at the loss-coupled block, one expert concentrates 91–95% of its dispatch weight on the `[CLS]` token alone, effectively becoming a dedicated CLS specialist, evidence that the router discovers content-dependent structure rather than an arbitrary partition of patches.

4.2.4 Downstream Transfer and Comparison with Prior Work

Standard fine-tuning is not sufficient to transfer MoE-pretrained representations: unfrozen routing can overwrite the patch-expert associations learned during pretraining, and the additional expert parameters increase overfitting risk on smaller downstream datasets. Three complementary adaptation strategies address this: Freeze Routing (FR), which preserves pretrained patch-expert assignments; Freeze Attention (FA); and Expert Dropout (ExD), which regularizes against over-reliance on any single expert. The combined recipe (FR+FA+ExD) improves 64-expert fine-tuning from 83.8% to 85.3% (a +1.5 point gain over vanilla MoE fine-tuning), exceeding the non-MoE baseline of 84.8%, and closes a 2.5–2.9 mIoU gap on ADE20K semantic segmentation, bringing MoE models to parity with or slightly ahead of their non-MoE counterparts.

Table 4.2 compares against published self-supervised methods on ImageNet-1K. Because Soft-MoE with a single slot per expert decouples parameter count from inference cost, each layer routes all tokens into only E expert slots rather than processing N tokens per expert, the 64-expert configuration uses *fewer* inference GFLOPs (13.86) than every ViT-B/16 comparator (17.45), despite having roughly 21.6× more parameters.

Table 4.2: Comparison with published self-supervised methods on ImageNet-1K (ViT-Base encoders). Lin. = linear-probe top-1, FT = fine-tuning top-1.

Method	Params	GFLOPs	Epochs	Lin.	FT
MAE He et al. (2022b)	86M	17.45	1600	68.0	83.6
BEiT Bao et al. (2021)	86M	17.45	800	56.7	83.0
SdAE Chen et al. (2022)	86M	17.45	300	—	84.1
MaskDistill Peng et al. (2022)	86M	17.45	300	76.2	84.8
MILAN Hou et al. (2022)	86M	17.45	400	79.9	85.4
MaskCLIP Dong et al. (2023)	86M	17.45	25	73.7	83.6
ExPLoRe (2-expert)	116M	11.93	300	79.6	84.1
ExPLoRe (64-expert)	1.86B	**13.86**	300	**80.6**	**85.3**

4.3 Summary

This work confirms three principles that carry forward into later chapters. First, **loss-coupling, not expert capacity, drives the improvement**: additional experts without a gradient pathway from the loss to the router are actively harmful, while dispatch-weight loss-coupling turns that same capacity into a meaningful gain. Second, **routing dynamics exhibit sharp transitions rather than smooth trade-offs**: entropy regularization separates stable specialization from catastrophic collapse, with the optimal operating point depending on expert count. Third, **transferring MoE representations downstream requires dedicated adaptation**, since naive fine-tuning discards the routing structure learned during pretraining.

More broadly, this chapter extends Phase 1's consistency principle to a new axis: just as explicit mechanisms for local-global and cross-scale consistency outperformed uniform treatment in Chapter 3, an explicit mechanism for patch-level adaptive weighting outperforms the uniform, global weighting used throughout prior multi-objective MIM work. That said, the experiment also exposes a limitation motivating the next chapter. The routing patterns discovered here are useful but not interpretable in any direct sense: we can observe, after training, that one expert specializes in the `[CLS]` token, but the framework provides no mechanism for a user to inspect, predict,

or control which patches an expert will emphasize before training completes, and the resulting representations inherit this opacity. Strong downstream performance does not by itself reveal what semantic factors are encoded, whether they are separable, or whether they could be manipulated independently. This gap between high performance and an interpretable, controllable structure motivates Phase 3's shift toward explicitly disentangled representations.

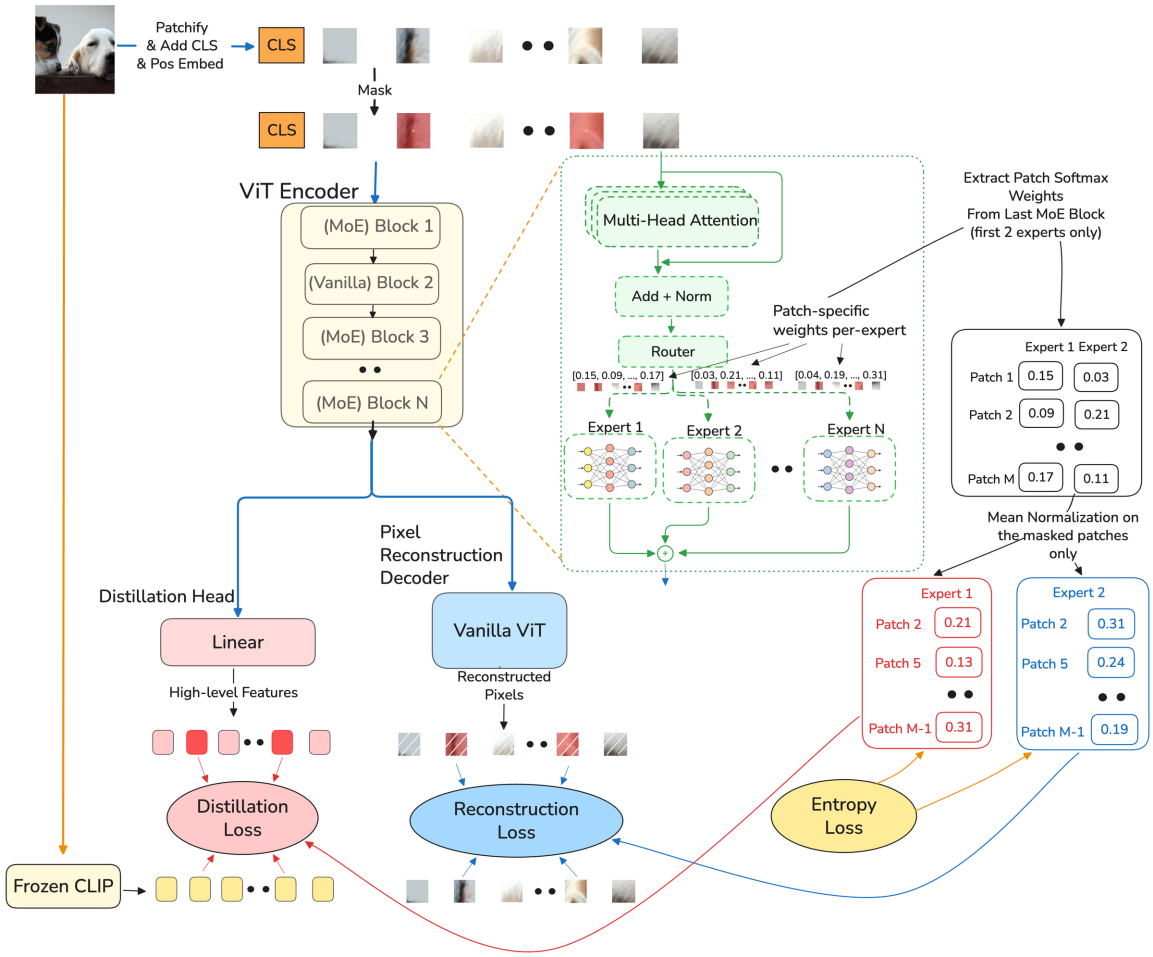


Figure 4.1: ExPLoRe framework overview. Soft Mixture-of-Experts (Soft-MoE) is integrated into the student encoder for patch-level adaptive loss weighting. The student encoder (ViT-Base with alternating MoE blocks) processes patches while a frozen CLIP teacher provides semantic targets. Soft-MoE dispatch weights serve as per-patch loss coefficients, with each expert weighting a different training objective. Loss-coupling, in which loss gradients flow through the dispatch weights back to the router, is the mechanism that enables learned, content-dependent specialization.

Chapter 5

Refinement through Disentanglement

The preceding chapters argued that structure should be declared rather than left to emerge, and tested that claim on scale and on the choice of learning signal. Both delivered stronger representations. Neither made the resulting representation any easier to read: a model can encode scale explicitly and still distribute object identity, pose, and background across the same directions in latent space. Declaring that structure exists is only half of what is needed. The other half is that the declared components be separable, so that one can be examined without dragging the others along with it.

This chapter takes up the second half. XTRA Tang and Qi (2026) adds factor tokens, dedicated learnable tokens that give distinct semantic factors somewhere of their own to live, and constrains them with a minimum volume constraint (MVC) that penalizes redundancy among them. The result improves disentanglement substantially over strong self-supervised baselines without giving up downstream accuracy, which is the evidence that interpretability and performance are not in tension. What it does not do is generate: factor tokens describe the structure already

present in a representation, and turning description into control is a separate problem that Chapter 6 returns to.

5.1 The Interpretability Gap

Chapter 3 and Chapter 4 achieved strong performance on diverse benchmarks, yet analyzing what these models actually learned revealed a limitation neither chapter set out to address: the representations remained largely opaque. While they clustered semantically similar images and transferred well to downstream tasks, semantic factors like object identity, pose, scale, color, lighting, and background were entangled within the latent space; changing one factor's value would unpredictably affect others. This entanglement limits what can be done with the representation. An attribute cannot be inspected without the others moving with it, which rules out attribute-level manipulation and makes it difficult to tell whether a model has learned a combination of factors or memorized their co-occurrence.

Disentangled representation learning has been extensively studied through VAE-based generative models such as FactorVAE Kim and Mnih (2018), which encourage independence among latent dimensions via modified training objectives. However, these approaches face challenges with high-resolution natural images, require careful hyperparameter tuning, and are primarily designed for generative rather than discriminative tasks. For discriminative vision transformers trained on complex natural images, achieving disentanglement without sacrificing performance remains challenging. Recent token-based approaches—register tokens Darcet et al. (2023), grouping tokens Xu et al. (2022), slot attention Seitzer et al. (2023)—demonstrate that additional learnable tokens can serve specialized purposes, but none directly address factor-level disentanglement in discriminative models. This gap motivated our factor token approach: augmenting vision transformers with dedicated tokens that explicitly capture distinct semantic factors, enforced by a constraint on the geometry of the factor space rather than on an assumed set of ground-truth factors.

There is a reason this gap has proven hard to close. Independence is a property one would like to supervise directly, but supervision requires knowing what the factors are, and for natural images nobody does. There is no label for "pose" that is separable from "viewpoint," no ground-truth enumeration of the factors that generated an ImageNet photograph. Methods that assume such an enumeration work on synthetic benchmarks and do not transfer. Any usable constraint must therefore act on a property of the representation that can be measured without naming the factors it is separating. The remainder of this chapter argues that redundancy is such a property.

5.2 XTRA

XTRA rests on a single requirement, and the architecture follows from it. If a representation is to be read one factor at a time, the factors must occupy positions that do not overlap: whatever is stored in one must not also be recoverable from another, or examining it in isolation means nothing. Two things are therefore needed. The representation needs designated places to put distinct factors, which is what the factor tokens of Section 5.2.1 provide, and it needs something to stop those places from collapsing into one another, which is what the constraint of Section 5.2.2 supplies. Neither is useful alone. Section 5.2.3 describes how the two are trained together.

5.2.1 XTRA Architecture

The core innovation is augmenting the standard vision transformer architecture with K learnable factor tokens, each designed to capture a distinct variation factor. Given an input image $\mathbf{X} \in \mathbb{R}^{H \times W \times 3}$, we follow standard ViT processing: divide into patches, linearly embed each patch, and add positional encodings. The key difference is prepending K factor tokens $\{\mathbf{f}_k\}_{k=1}^{K}$ to the sequence:

$$\mathbf{z}_0 = [\mathbf{f}_1, \ldots, \mathbf{f}_K, \mathbf{p}_1, \ldots, \mathbf{p}_N] + \mathbf{E}_{\text{pos}}$$

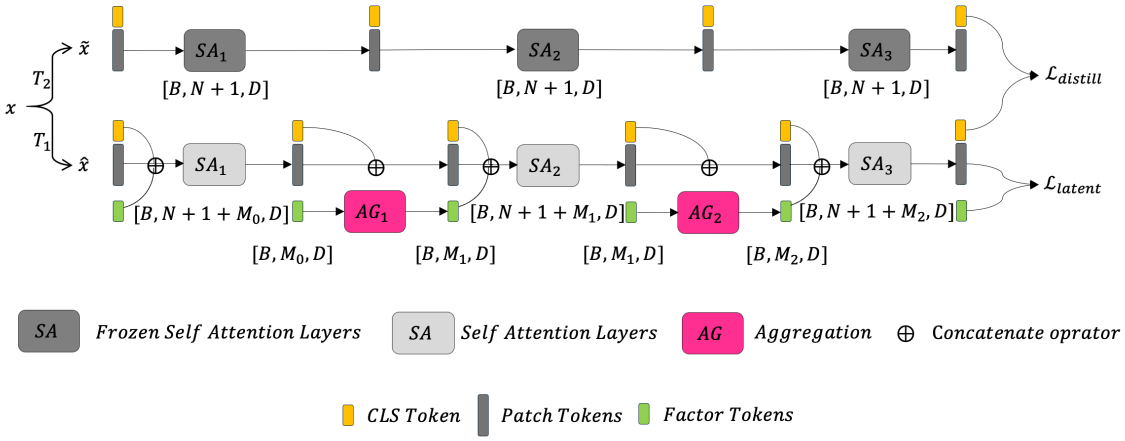


Figure 5.1: Illustration of XTRA built upon the dual-stream self-knowledge distillation network. Top: teacher network. Bottom: The multi-stage aggregation student network.

where $\mathbf{f}_k \in \mathbb{R}^D$ are learnable parameters initialized randomly. Unlike the standard class token that aggregates information for classification, factor tokens are explicitly designed to capture orthogonal variation factors through the training objective.

The sequence processes through L transformer layers. Critically, factor tokens interact with patch tokens and each other through self-attention, allowing them to gather information about their designated factors from the visual content:

$$\mathbf{z}_\ell = \text{TransformerBlock}_\ell(\mathbf{z}_{\ell-1})$$

After the final layer, we extract the factor token representations $\{\mathbf{z}_L^{(k)}\}_{k=1}^K$ which form the disentangled representation. For downstream tasks like classification, we can either use only the factor tokens (concatenated or pooled), combine them with a standard class token, or use the full token sequence depending on task requirements.

Two design parameters govern the factor tokens. The first is their number, K, which trades expressiveness against separability: too few factors force distinct semantic properties to share a token, while too many leave the constraint trying to

hold apart directions that carry no distinct content. We use $K = 8$ to $K = 12$ for natural images. The second is their dimensionality, which we set equal to the model dimension D, so that factor and patch tokens inhabit the same space and the mixing model of Eq. 5.1 is well posed without an intermediate projection. Factor tokens are initialized randomly; seeding them with embeddings chosen to encourage particular factors is possible when such factors are known in advance, though none of the experiments reported here rely on it.

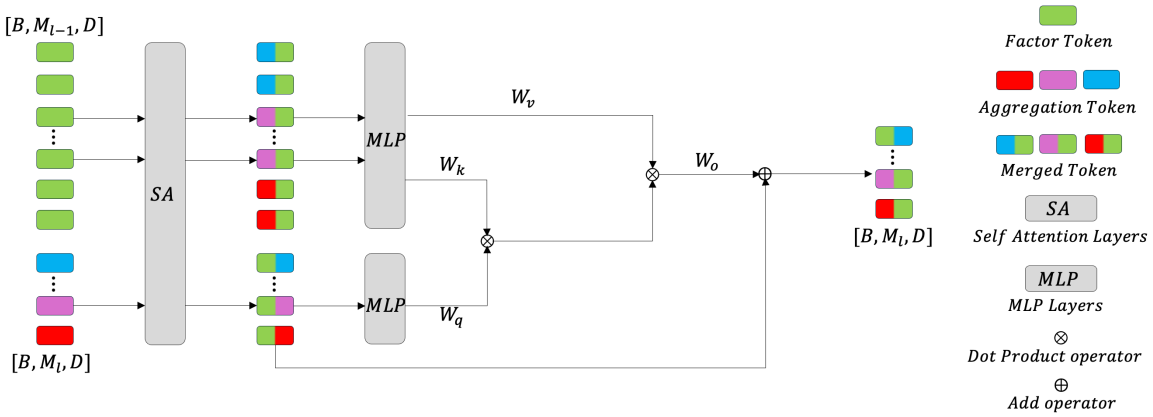


Figure 5.2: Illustration of a 2-stage aggregation of factor tokens

5.2.2 Minimum Volume Constraint

Factor tokens supply slots. They do not, on their own, supply any reason for the slots to differ from one another, and a model free to fill them as it likes will happily place the same information in several. The constraint developed here is what makes the slots distinct.

Begin with what the factor tokens are meant to be. The goal is not that they store high-norm or otherwise anomalous content, as the register tokens of Darcet et al. (2023) do, but that they act as a basis from which the patch tokens can be reconstructed. Given N patch tokens $\{\mathbf{p}_i\}_{i=1}^{N}$ and K factor tokens $\{\mathbf{f}_k\}_{k=1}^{K}$, we ask

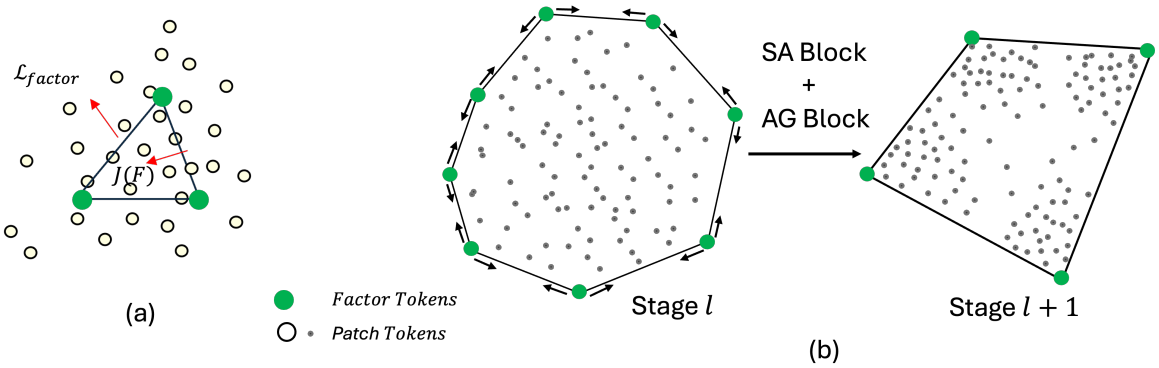


Figure 5.3: (a) A geometric illustration of the two loss terms within the latent loss (Eq. 5.4) where the minimum volume constraint, $J(F)$, serves as the internal force pointing inward and the patch reconstruction constraint, $\mathcal{L}_{factor}$, serves as the external force pointing outward. (b) Illustration of how the factor tokens and patch tokens evolve across two stages of aggregations.

that each patch be expressible as a mixture of the factors,

$$\mathbf{p}_i = F \cdot \mathbf{w}_i, \qquad F = [\mathbf{f}_1, \cdots, \mathbf{f}_K] \tag{5.1}$$

where $\mathbf{w}_i$ weights each factor's contribution to patch i. This is the linear mixing model standard in spectral unmixing (Miao and Qi, 2007), with factor tokens in the role of endmembers and patch tokens in the role of observed mixtures.

Posed this way, the requirement on the factors becomes concrete. A basis whose elements carry overlapping information is a redundant basis: two factors that encode the same thing describe a smaller space between them than two genuinely distinct factors would. Redundancy, unlike independence, can be measured without knowing what the factors are, which is what makes it a usable target given the difficulty raised in Section 5.1.

The measurement is geometric. The volume of the parallelepiped spanned by K vectors in $\mathbb{R}^D$ (with $K \leq D$) is

$$\mathrm{Vol}(\mathbf{f}^{(1)}, \ldots, \mathbf{f}^{(K)}) = \sqrt{\det(\mathbf{F}^T\mathbf{F})}, \tag{5.2}$$

and the Gram matrix $\mathbf{G} = \mathbf{F}^T\mathbf{F} \in \mathbb{R}^{K\times K}$, with entries $G_{ij} = \mathbf{f}^{(i)T}\mathbf{f}^{(j)}$, records exactly the quantities of interest. Its off-diagonal entries are the pairwise correlations between factors; its diagonal entries are their squared norms. Driving the off-diagonals toward zero *is* what separating the factors means in this representation, and it is the operation the constraint performs.

For factors of fixed length, this has a direct reading in terms of volume. When the factors are mutually orthogonal, $\mathbf{G}$ is diagonal, and the volume is at its maximum; as they become correlated, $\det(\mathbf{G})$ falls, reaching zero when two factors coincide. Minimizing redundancy among the factors is therefore the same thing as maximizing the volume they enclose.

The name *minimum volume constraint* is inherited from the unmixing literature and refers to a different volume. There the objective is to find the smallest simplex enclosing the observed mixtures, so that the endmembers sit tightly around the data rather than far outside it. The two formulations agree: the simplex is what shrinks, and it shrinks by holding the factors to unit length while spreading them apart, which is a maximum-volume condition on the factors themselves. We keep the established name and state the objective in the form we optimize.

That objective penalizes departure of the Gram matrix from the identity,

$$J(F) = \|F^TF - I\|_F^2, \tag{5.3}$$

which carries both requirements in a single term. The off-diagonal entries impose orthogonality, removing redundancy. The diagonal entries hold each factor at unit norm, which matters because without it the reconstruction term in Eq. 5.4 could be satisfied by inflating the factors rather than by arranging them well. This is the sense in which $J(F)$ acts as the inward force in Figure 5.3: it caps the extent of the basis while the reconstruction term pushes it outward to cover the data. Formulating the penalty on $\mathbf{G}$ rather than on $\det(\mathbf{G})$ directly also avoids the numerical fragility of

differentiating a determinant that approaches zero precisely in the failure case the constraint is meant to prevent.

The latent loss combines the two forces:

$$\mathcal{L}_{\text{latent}} = \lambda_{\text{factor}} \cdot \mathcal{L}_{\text{factor}} + \lambda_{\text{volume}} \cdot J(F) \tag{5.4}$$

$$\mathcal{L}_{\text{factor}} = \frac{1}{2} \log \left(\sum_{i=1}^{N} \|\mathbf{p}_i - F \cdot \mathbf{w}_i\|^2 \right) \tag{5.5}$$

where λ_{volume} controls the strength of the volume penalty. Neither term is sufficient alone. Reconstruction without the volume penalty admits a redundant basis that covers the data with correlated factors; the volume penalty without reconstruction admits a well-separated basis that represents nothing. Table 5.2 reports what happens when the volume penalty is removed.

5.2.3 Training Objective and Methodology

XTRA is a dual-stream neural network, consisting of a standard vision transformer stream for all the patch tokens and a multi-stage aggregation stream for the factor tokens. Specifically, rather than concatenating only one trainable `[CLS]` token with the patch tokens, the M trainable factor tokens are also concatenated with the patch tokens. These trainable tokens are then fed to the designed network that outputs the learned `[CLS]` token, patch tokens, and M factor tokens, after L aggregation stages. Following the standard self-knowledge distillation framework, given the image x, we first apply random data augmentations to generate distinct views. For clarity, we consider two views, i.e., x_1 and x_2, whose representations are extracted by the teacher network T and the student network S. So, $\left[\hat{c}, \hat{f}, \hat{p}\right] = T(x_1)$ and $\left[\tilde{c}, \tilde{f}, \tilde{p}\right] = S(x_2)$, respectively. Then, the `[CLS]` tokens are further processed using projection heads, i.e., $\hat{h}^c = proj\,(\hat{c})$ and $\tilde{h}^c = proj\,(\tilde{c})$.

In XTRA, we select the asymmetric contrastive loss to measure the similarity between the [CLS] tokens output from the teacher and the student networks, representing the distillation loss, $\mathcal{L}_{\text{distill}}$, and is defined as

$$\mathcal{L}_{distill} = \mathcal{L}_{\hat{h}^c \leftrightarrow \tilde{h}^c} = \mathcal{L}_{\hat{h}^c \to \tilde{h}^c} + \mathcal{L}_{\tilde{h}^c \to \hat{h}^c} \tag{5.6}$$

where the two asymmetric contrastive losses are defined as

$$\begin{aligned} \mathcal{L}_{\hat{h}^c \to \tilde{h}^c} &= -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp\left(\hat{h}_i^c \cdot \tilde{h}_i^c / \tau\right)}{\sum_{j=1}^{B} \exp\left(\hat{h}_i^c \cdot \tilde{h}_i^c / \tau\right)} \\ \mathcal{L}_{\tilde{h}^c \to \hat{h}^c} &= -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp\left(\tilde{h}_i^c \cdot \hat{h}_i^c / \tau\right)}{\sum_{j=1}^{B} \exp\left(\tilde{h}_i^c \cdot \hat{h}_i^c / \tau\right)} \end{aligned} \tag{5.7}$$

Here, B is the batch size. The [CLS] token is often adopted to encode the global context, which could be a good representation for global semantic information. However, it may be less representative of factors controlling different aspects of an image, such as foreground/background, object position/rotation, object properties, etc. To enhance the representational capacity for explainability and disentanglement, we introduce factor tokens, which can complement existing representations to capture finer, more localized scales. Specifically, we characterize the properties of the latent space spanned by the factor tokens and examine the relationship between factor tokens and patches.

The complete training objective combines the distillation task loss with MVC:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{distill}} \cdot \mathcal{L}_{\text{distill}} + \lambda_{\text{factor}} \cdot \mathcal{L}_{\text{factor}} + \lambda_{\text{volume}} \cdot J(F) \tag{5.8}$$

where λ_{distill}, λ_{factor}, and λ_{volume} are hyperparameters that control the trade-off among the different loss terms.

The weight λ_{volume} on the volume penalty is the one term that requires care, because it sets the balance between the two forces of Eq. 5.4. Below roughly 0.01 the

penalty is too weak to separate the factors and the reconstruction term dominates, producing a basis that covers the data with correlated directions. Above roughly 1.0 the constraint begins to cost downstream accuracy, holding the factors apart at the expense of what they encode. We use $\lambda_{\text{volume}} \in [0.1, 0.5]$, with the exact value depending on task and dataset.

The penalty is inexpensive to compute. Forming $\mathbf{G} = \mathbf{F}^T\mathbf{F}$ costs $O(K^2D)$ with $K \leq 16$, and the Frobenius norm of $\mathbf{G} - \mathbf{I}$ requires no factorization, which is the practical reason for preferring it to a determinant that must be differentiated as it approaches zero. We apply the penalty at every training step, over the factor tokens taken from the final transformer layer. Optimization is otherwise standard: AdamW, a cosine learning rate schedule, random crops, flips, and color jittering for augmentation, and distributed training across multiple GPUs.

5.2.4 Experimental Validation

We evaluate disentanglement quality, which is our primary contribution, then show that this disentanglement *simultaneously* improves representation quality across multiple tasks.

Disentanglement Quality

Since disentangled representation learning by explicit regularization is the main claim of XTRA, in this set of experiments, we evaluate the degree of disentanglement of the learned representation. Given no ground truth, we follow Wang et al. (2024b) and adopt an unsupervised disentanglement metric SEPIN@k (Do and Tran, 2021). SEPIN@k measures how each token $\{\mathbf{p}_i\}$ is disentangled from others $\{\mathbf{p}_{\neq i}\}$ by computing their conditional mutual information with the top k features.

As shown in Table 5.1, the representation from XTRA exhibits significantly better disentanglement than DINO v2 and its variant in all top-k dimensions. Since the

Table 5.1: Representation disentanglement score with SEPIN@k on ImageNet-1k, where k denotes the top-k dimensions (higher is better).

	SEPIN@1	SEPIN@10	SEPIN@100	SEPIN@all
DINO v2	0.47 ± 0.03	0.39 ± 0.02	0.28 ± 0.02	0.11 ± 0.01
DINO v2 + Register	0.42 ± 0.02	0.35 ± 0.03	0.25 ± 0.01	0.13 ± 0.01
XTRA	3.95 ± 0.12	3.02 ± 0.09	1.54 ± 0.06	0.16 ± 0.04

learned features also contain noisy components, the all-dimensional ($k = 768$) results are similar across methods, with XTRA still maintaining a slight advantage.

In Tang and Qi (2026), we further show the representation SEPIN@k score at the different aggregation stages, where the first two stages are the representation after aggregation, and the last stage is the output representation. For comparison purposes, we also use DINO v2 and DINO v2-Reg, both of which have four self-attention blocks at each stage. The results again demonstrate that factor token aggregation substantially enhances the disentanglement of representations.

Practical Benefits of Part-Level Disentanglement

We validate that XTRA's part-level disentanglement provides practical benefits through part segmentation. We evaluate part segmentation on PartImageNet (He et al., 2022a), which contains pixel-level part annotations for 158 ImageNet categories. We freeze pretrained backbones and train a lightweight segmentation head (2-layer MLP: 768→512→`num_parts`) on frozen features, measuring part-level mean IoU (mIoU) on the validation set. Table 5.2 shows XTRA achieves 46.8% mIoU vs. DINOv2's 42.3% (+4.5 mIoU, $p < 0.01$). Critically, Table 5.3 shows largest improvements on *articulated parts*: legs +6.5%, tail +7.4%, ears +6.1%.

Factor-Part Alignment. We also compute the overlap between XTRA's factor token attention and ground-truth semantic parts, finding an average overlap of 0.81 ± 0.06. For quadrupeds: Factor 0→head (0.82 overlap), Factor 1→body (0.88), Factor 2→legs (0.79), Factor 3→tail (0.81). This directly shows that factors have learned part-level decomposition during pretraining.

Table 5.2: Part segmentation on PartImageNet. XTRA achieves superior part-level mIoU, with largest improvements on articulated parts.

Method	Backbone	Part mIoU (%)	vs. DINOv2
DINOv2	ViT-B/16	42.3 ± 0.4	baseline
DINOv2 + Register	ViT-B/16	43.1 ± 0.3	+0.8
XTRA (no MVC)	ViT-B/16	43.5 ± 0.5	+1.2
XTRA (full)	ViT-B/16	$\mathbf{46.8 \pm 0.3}$	**+4.5**

Table 5.3: Per-part breakdown for the "Quadruped" category.

Method	Head	Body	Leg	Tail	Ear	Mean
DINOv2	51.2	68.4	38.7	34.2	42.8	47.1
XTRA	56.8	71.3	45.2	41.6	48.9	52.8
Δ	+5.6	+2.9	+6.5	+7.4	+6.1	+5.7

Representation Quality with Pretrained Teacher

KNN & Linear Probing. Following standard self-supervised evaluation protocols, we evaluate XTRA's representations on ImageNet-1K using KNN and linear-probe accuracy, as shown in Table 5.4. We observe that XTRA outperforms all prior ImageNet-1K pre-training methods by 2.1% in KNN and 1.5% in linear probing. XTRA also outperforms every DINO model, including DINO v2 and its variant with register. It demonstrates that XTRA, using a foundation model as a teacher, can generate better representations, with a lightweight trainable student network and extra token regularization.

We further investigate the performance of the learned representation without the strong pre-trained teacher network. For fair comparison, we use the same backbone, ViT-Base, and pre-train both DINO v2 and XTRA on the same dataset, ImageNet-1K. The results are reported in Table 5.5. We observe that even without a pre-trained foundation model as a teacher, XTRA maintains its superior performance compared to these large-scale foundation models.

Table 5.4: Evaluation of representation from pre-trained model in KNN and Linear Probing (%).

	Backbone	Dataset	Epochs	KNN	Linear
MoCo-v3	ViT-B	IN-1K	1200	51.2	76.3
MAE	ViT-B	IN-1K	800	54.75	71.8
BEiT	ViT-B	IN-1K	800	49.06	56.7
iBOT	ViT-B	IN-1K	1600	72.9	82.3
DINO v1	ViT-B	IN-1K	300	76.1	78.2
OpenCLIP	ViT-G	IN-1K	800	75.2	78.2
OpenCLIP + Reg	ViT-G	IN-1K	800	75.8	78.1
DINO v2	ViT-G	LVD-142M	-	82.1	84.5
DINO v2 + Reg	ViT-G	LVD-142M	-	82.0	83.6
XTRA	ViT-B	IN-1K	300	**84.2**	**86.0**

Table 5.5: Evaluation of representation without pre-trained teacher network in KNN and Linear Probing on IN-1K(%).

	Teacher	Backbone	KNN	Linear
DINO v2	None (from scratch)	ViT-Base	76.9	80.1
DINO v2 + Register	None (from scratch)	ViT-Base	77.3	82.1
XTRA	DINOv2 (None, from scratch)	ViT-Base	**81.9**	**83.8**
XTRA	DINOv2 (LVD142M, pretrained)	ViT-Base	**84.2**	**86.0**

Downstream Tasks We evaluate XTRA's generality by fine-tuning on three downstream benchmarks—ImageNet-1K classification, ADE20K semantic segmentation, and COCO2017 object detection—each for 100 epochs. Table 5.6 summarizes Top-1 accuracy for classification, mIoU for segmentation, and AP_{box} for detection. XTRA surpasses state-of-the-art methods on ImageNet classification, ADE20K segmentation, and COCO2017 object detection, demonstrating the robustness of its learned representations across diverse tasks.

Additional ablation studies are reported in the original publication Tang and Qi (2026).

Table 5.6: Evaluation of representation from pre-trained model with different downstream tasks (%).

	Classification (Top-1)	Segmentation (mIoU)	Detection (AP box)
Backbone	ImageNet 1K	ADE20K	COCO2017
MoCo-v3	83.1	47.3	47.9
MAE	83.6	48.1	50.3
BEiT	83.2	47.1	49.8
iBOT	84.0	50.0	48.2
DINO v1	82.8	51.3	46.8
DINO v2	85.8	54.4	51.2
DINO v2 + Register	85.6	54.2	50.5
XTRA	**85.9**	**55.1**	**52.1**

5.2.5 Factor Token Visualization

We consider the factor tokens as "mixtures" of semantic contents in the scene. By incorporating the minimum volume constraint and the consistency constraint between the factor tokens and patch tokens, we generate attention maps with much finer detail while preserving semantic consistency—factor tokens capture different parts of the object, including some with high norms Darcet et al. (2023), as shown above. Here we present additional examples. See Fig. 5.5 and Fig. 5.6. In Fig. 5.5, we show results on different animals, illustrating the capability of factor tokens to capture semantic parts, including a factor token with a high norm. In Fig. 5.6, we show results on non-animal objects, where capturing the whole object proves easier than disentangling it into parts. We think this may be because these objects are difficult to disentangle into part-wise properties, a limitation examined further below.

Failure Case Analysis

In Fig.5.6, airplanes and ambulances exhibit less clear part decomposition, which constitutes a failure case for disentanglement performance. We think some potential reasons cause the failure case: **(1) Rigid objects with uniform appearance:** Unlike animals with distinct part textures (fur patterns, facial features), vehicles have more uniform surfaces. The semantic "parts" (wings, fuselage for

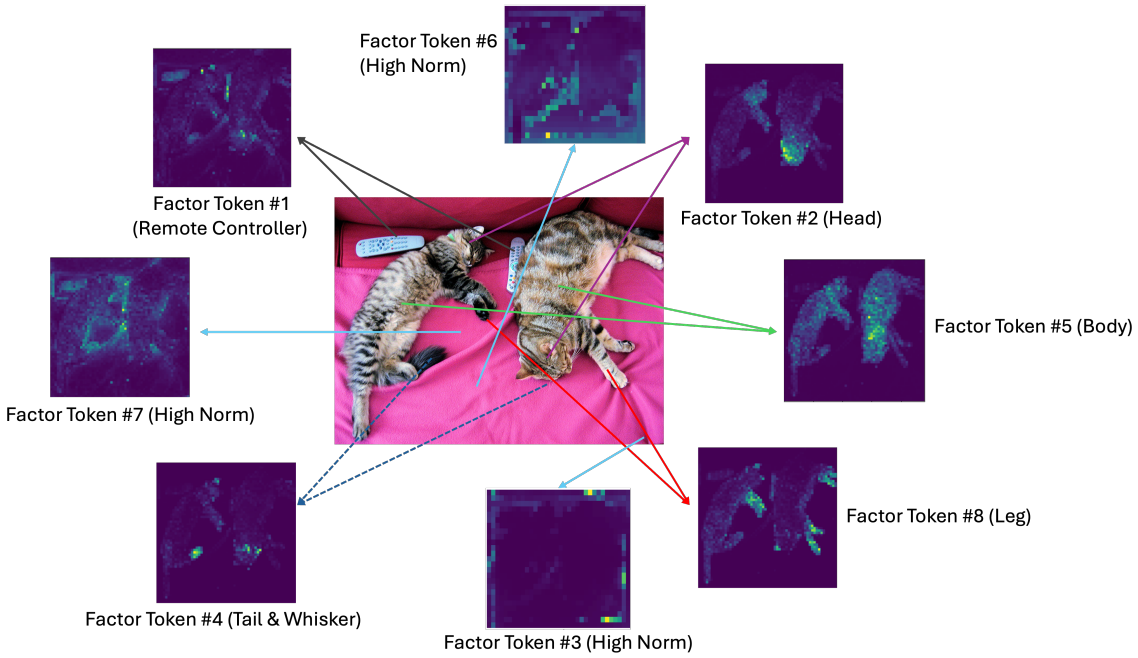


Figure 5.4: XTRA enables disentangled attention maps pertaining to consistent parts across multiple objects in the scene.

planes; body, wheels for ambulances) are less distinct in the learned feature space.**(2) Training data bias:** ImageNet-1K contains more animals than vehicles, potentially biasing part discovery toward biological structures that appear more frequently during training. **(3) Semantic ambiguity:** For vehicles, "parts" (hood, door, wheel) may be less distinct in feature space than animal parts (head, leg, tail) because vehicles have more uniform color/texture (e.g., all parts are painted the same color), less deformable structure (rigid vs. articulated), and less consistent part arrangement (cars vary more in design than animal body plans).

5.3 Summary

This chapter demonstrated that factor tokens augmented with a Minimum Volume Constraint successfully impose an interpretable, disentangled structure on vision

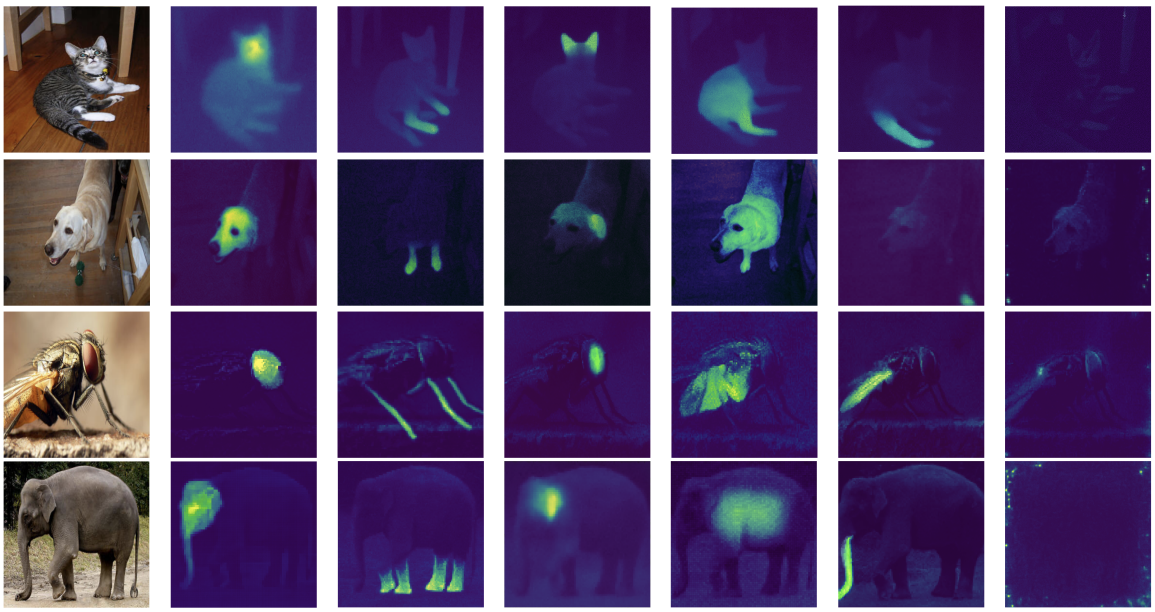


Figure 5.5: Visualization of Factor Tokens

transformer representations without sacrificing downstream task performance. Using an unsupervised disentanglement metric (SEPIN@k) on ImageNet-1K, XTRA achieves substantially better disentanglement than DINOv2 while simultaneously improving representation quality (kNN and linear-probe accuracy) and downstream performance on classification, segmentation, and detection. Critically, this disentanglement is not merely a favorable metric: part-segmentation experiments on PartImageNet confirm that individual factor tokens correspond to actual object parts, head, body, legs, tail, ears, with the largest gains on the articulated parts whose boundaries are hardest to learn implicitly. This validates that interpretability and performance are not mutually exclusive when disentanglement is designed correctly into the architecture through dedicated tokens and geometric constraints.

The success of factor tokens addresses the limitation identified in Chapter 4: representations are no longer black boxes but have explicit structure, enabling both understanding and control. This work focused on discriminative tasks, classification and recognition, where disentangled representations aid interpretability without

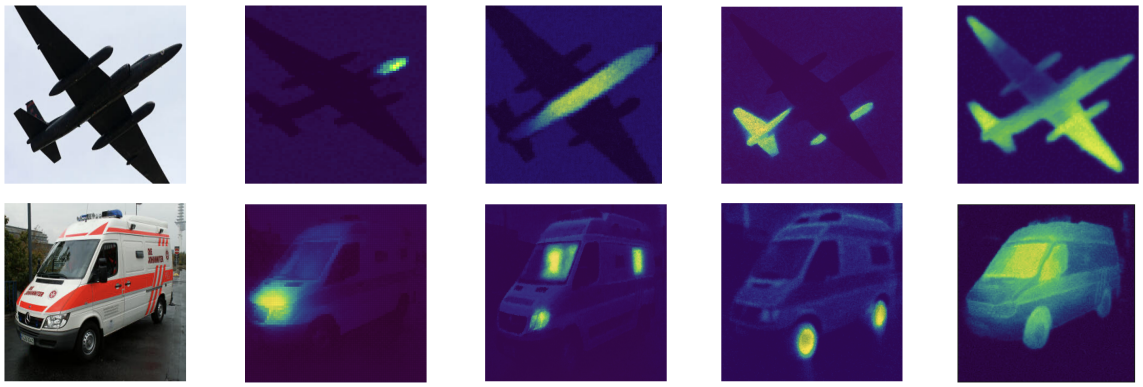


Figure 5.6: Visualization of Factor Tokens (Other Objects)

requiring generation. Whether the same disentangled structure could support controllable generation, or whether a similar mechanism could benefit other domains such as remote sensing, are open questions this dissertation returns to in Chapter 6's discussion of future directions.

Chapter 6

Discussion and Future Directions

The preceding three chapters each made a case and each left something open. We showed in Phase 1 that heterogeneous content needs explicit, non-uniform treatment, across local versus global structure and across scale, where uniform processing falls short. Phase 2 carried the same principle onto a different axis entirely: not what scale a patch was captured at, but which learning signal it should be trained with. Phase 3 asked a different question, and found that imposing explicit, interpretable structure costs nothing in task performance. Each contribution is complete and independently validated. None closed every question it raised. We read those open questions together here, and follow them to the two directions they point toward.

6.1 Limitations of the Completed Work

Phase 1's two contributions each have a documented scope. Global-local balanced contrastive learning Tang et al. (2023b) operates at a fixed spatial resolution; when evaluated on datasets that mix different ground sampling distances, its performance degrades substantially, indicating that representations learned at one scale do not automatically transfer to another. Cross-Scale MAE Tang et al. (2023a) addresses scale variation directly, but at roughly 3× the computational cost of single-scale MAE, and large extrapolations beyond the training scale range still show degraded

performance: the representations it learns are more scale-robust, not scale-invariant, and scale is not explicitly disentangled from content within them.

Phase 2's ExPLoRe Georgiou et al. (2026) mechanism was validated with a frozen CLIP teacher and up to 64 experts, a scale bounded by available GPU memory rather than by the method itself. The original paper's own discussion identifies natural extensions it did not pursue: a dedicated loss-weighting router separate from the router that handles feature computation, multiple slots per expert, and learned aggregation of dispatch weights across layers. Whether the mechanism's benefits persist at substantially larger scale, or with teachers other than CLIP, remains untested.

Phase 3's XTRA Tang and Qi (2026) framework has two limitations acknowledged directly in the original publication. First, disentanglement quality correlates with the extent to which the teacher's feature space assigns within-class variance to an object category; rigid, texturally uniform objects such as vehicles yield visibly weaker part decompositions than articulated ones such as animals. Second, and more consequential for what follows, the framework is descriptive rather than compositional: factor tokens characterize the structure already present in a pretrained representation, but the model was never asked to generate anything from that structure. Interpreting a representation and controlling what a model produces are different problems, and XTRA addresses only the first.

Taken together, these limitations point in two directions. The gap between descriptive and compositional disentanglement points toward controllable generation. The combination of unresolved scale-content entanglement (Phase 1), an adaptive-routing mechanism validated only on natural images (Phase 2), and remote sensing's original motivating characteristics, extreme scale variation, multiple sensors, multiple spectral bands, points back toward the domain this dissertation began with.

6.2 Controllable Generation via Disentangled Latent Spaces

Current diffusion models control attributes through indirect mechanisms: text prompts provide coarse semantic guidance, spatial conditioning, such as ControlNet, enables layout control, and post hoc latent manipulation offers limited, often entangled, editability. None of these gives a user the kind of fine-grained, independent control that an explicitly disentangled latent space would provide in principle. Whether factor tokens can supply that latent space is a question XTRA's own limitations leave open rather than answer.

One direction, discussed here without commitment to a specific architecture, would augment a Diffusion Transformer Peebles and Xie (2023) with K learnable factor tokens maintained under a minimum volume constraint during training, so that a user could specify desired factor values, interpolate between them for smooth transitions, or compose factors drawn from different sources. Whether this works is not obvious in advance. Diffusion training optimizes sample fidelity across an entire data distribution, a different objective than the discriminative task XTRA's factor tokens were designed for, and it is unclear whether an MVC-style constraint tuned for classification-oriented disentanglement would transfer to a generative latent space without degrading sample quality, or whether it should apply uniformly throughout the diffusion process rather than at a restricted subset of steps. It is equally unclear whether the factors a discriminative teacher makes easy to disentangle, the part-level structure XTRA discovers from DINOv2, are the same factors most useful for controllable generation, since attributes such as lighting or viewpoint may matter more for synthesis than for recognition. A staged approach of the kind XTRA itself used, establishing a strong generative backbone first and only then studying how much disentanglement pressure it tolerates before quality suffers, seems like a reasonable way to approach these questions, though it is one approach among several one might take.

6.3 Efficient, Multi-Scale Remote Sensing Foundation Models

ExPLoRe's own discussion, not this dissertation's speculation, identifies the natural next step for its mechanism: applying per-patch adaptive routing at a larger scale, with a dedicated loss-weighting router, and to data beyond the natural images it was validated on. Remote sensing is a natural setting in which to pursue this, both because it motivated Phase 1 and because it combines exactly the kind of heterogeneity, across sensors, spectral bands, and the extreme ground-sampling-distance variation documented in Chapter 3, that a patch-adaptive mechanism is suited to address.

Combining Cross-Scale MAE's multi-scale processing with ExPLoRe's adaptive routing raises questions that neither paper, on its own, answers. It is not obvious whether experts would specialize along the axis that matters most for this domain, scale or ground sampling distance, in the way one expert specialized toward the `[CLS]` token on natural images, or whether the routing signal would need to be redesigned for scale-specific specialization to emerge at all. Experts might instead organize around spectral characteristics (RGB versus multispectral versus SAR imagery), geographic content, or downstream task, and which of these organizing principles would dominate is an empirical question rather than a design choice that can be assumed in advance. Nor is it clear whether the computational overhead Cross-Scale MAE already carries, roughly 3× single-scale MAE, would compound with or be offset by the efficiency Soft-MoE Puigcerver et al. (2024) provides through sparse expert activation; the two mechanisms were never evaluated together, and there is no principled reason to assume their costs and benefits simply add.

6.4 Summary

Phases 1 through 3 constitute this dissertation's completed contribution: domain-specific foundations in remote sensing, a demonstration that the underlying consistency principle generalizes to natural images, and a demonstration that interpretable structure and strong performance are compatible rather than competing goals. The limitations discussed above are not shortcomings held back until this chapter; each was identified in the chapter or publication in which the relevant method was introduced. What this chapter has done is to read them together across phases to see where they point. Controllable generation and efficient multi-scale foundation models for remote sensing are what they point toward, not commitments this dissertation makes on their behalf. Chapter 7 closes the dissertation by returning to what Phases 1 through 3 established.

Chapter 7

Conclusion

This dissertation began with a question raised in Chapter 1: can a representation be both high-performing and interpretable, or do these properties trade off against one another? Three contributions, across two domains, argue for the same answer. They are compatible, provided the structure a representation needs is built into its architecture and training objective rather than left for scale or capacity to discover on its own.

7.1 Summary of Contributions

We began in remote sensing because that is where the problem is impossible to ignore. Remote sensing imagery is heterogeneous in ways natural images typically are not: a fixed patch size spans wildly different physical extents depending on ground sampling distance, and dense prediction tasks demand both global semantic understanding and pixel-level spatial precision. Global-local balanced contrastive learning Tang et al. (2023b), published at WACV 2023, addressed the second form of heterogeneity with an explicit local-consistency loss operating alongside global contrastive objectives. Cross-Scale MAE Tang et al. (2023a), published at NeurIPS 2023, addressed the first with explicit cross-scale consistency and reconstruction objectives, later validated

on a genuinely difficult downstream task, construction-phase classification on multi-sensor imagery Georgiou et al. (2024) presented at IGARSS 2024. Both contributions outperformed methods that treat their respective forms of heterogeneity uniformly, establishing the principle that Phases 2 and 3 would go on to test in different settings.

We then asked whether the principle was a fact about satellites. Multi-objective masked image modeling combines complementary learning signals, token distillation, CLS alignment, and pixel reconstruction, but the field had settled on weighting them with a single global scalar, applied uniformly to every patch regardless of content. ExPLoRe Georgiou et al. (2026), accepted to ECCV 2026, repurposed Mixture-of-Experts routing to weight these objectives adaptively per patch, and isolated why this worked through a direct ablation: extra experts without a gradient pathway from the loss back to the router were actively harmful, while coupling that pathway improved both representation quality and downstream transfer. The underlying principle generalized beyond the domain, and beyond the specific form of heterogeneity on which it was established.

Phase 3 asked something else of the representation. Strong benchmark performance, on remote sensing imagery and on natural images alike, said nothing about whether a representation's semantic factors were separable or inspectable. XTRA Tang and Qi (2026), accepted to ICML 2026, introduced factor tokens constrained by a Minimum Volume Constraint to explicitly disentangle a vision transformer's representation into interpretable parts. On PartImageNet, the volume constraint accounted for the majority of XTRA's improvement over DINOv2 in part-level segmentation (+4.5 mIoU with the constraint versus +1.2 without it, all else held fixed), showing that the disentanglement was doing real work rather than incidentally correlating with it. Interpretability and performance improved together, but only because the disentanglement was designed in rather than hoped for.

7.2 The Case for Explicit Structure

Taken separately, these contributions solve different problems in different settings. Taken together, they make two claims.

The first is that the structure a representation needs should be declared rather than left to emerge. Phase 1 supports it by comparison: methods that treat local-global structure and cross-scale structure uniformly are outperformed by methods that handle them explicitly. Phases 2 and 3 support it more directly, each through an ablation that changes only the mechanism at issue. Removing the gradient pathway from loss to router made added expert capacity actively harmful rather than merely useless; removing the volume constraint reduced XTRA's gain over DINOv2 from +4.5 to +1.2 mIoU. In neither case did the property in question survive the removal of the mechanism that declared it.

The second claim is that declaring structure is not sufficient on its own. A model can encode scale explicitly, and route learning signals per patch, and still produce a representation in which no individual factor can be examined without the others moving with it. Phase 3 addresses this directly: factor tokens give distinct semantic factors somewhere of their own to live, and the volume constraint keeps those places from collapsing into one another. Explicit tokens gave the factors a place to live; the constraint kept them apart. That the same change improved both disentanglement and downstream accuracy is what makes the two properties compatible rather than competing.

Neither claim extends automatically beyond the settings tested here, and neither should be read as asserting that architectural priors always beat scale. Larger models trained on more data have, in other contexts, found structure that smaller and more constrained ones could not. What this dissertation shows is narrower: in these settings, spanning two domains and two distinct target properties, the explicit mechanism was not merely sufficient but necessary, and the cost of building it in was small relative to the benefit.

7.3 Closing Remarks

Chapter 6 discusses each contribution's limitations in detail, together with the two directions those limitations point toward: controllable generation, and foundation models for remote sensing. Both would extend the argument made here to settings this dissertation did not test. One replaces a discriminative objective with a generative one. The other returns to a domain that combines, in a single problem, every form of heterogeneity the three phases addressed separately.

The work in this dissertation was developed in response to specific, sometimes narrow problems: an aerial imagery segmentation benchmark, a multi-scale remote sensing dataset, a multi-objective pretraining recipe, and a disentanglement metric. None were designed from the outset as demonstrations of a general principle. That the same principle, explicit structure over implicit discovery, was what made each of them work is something we arrived at, not something we set out to prove. Whether the same principle continues to hold as these methods are extended to generation and to new domains is a question for the work Chapter 6 discusses, not one this dissertation can answer in advance.

Bibliography

Anciukevicius, T., Lampert, C. H., and Henderson, P. (2020). Object-centric image generation with factored depths, locations, and appearances. 17

Bao, H., Dong, L., Piao, S., and Wei, F. (2021). Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*. 2, 14, 47, 53

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, D. (2018). Mutual information neural estimation. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 531–540. PMLR. 26

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828. 2, 4

Biza, O., van Steenkiste, S., Sajjadi, M. S. M., Elsayed, G. F., Mahendran, A., and Kipf, T. (2023). Invariant slot attention: Object discovery with slot-centric reference frames. 17

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901. 11

Burgess, C. P., Matthey, L., Watters, N., Kabra, R., Higgins, I., Botvinick, M., and Lerchner, A. (2019). Monet: Unsupervised scene decomposition and representation. 17

Burt, P. J. and Adelson, E. H. (1987). The laplacian pyramid as a compact image code. In *Readings in Computer Vision*, pages 671–679. Elsevier. 12

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924. 25

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660. 9, 16, 35, 36, 47

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848. 13

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020a). A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR. 1, 20, 22

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020b). A simple framework for contrastive learning of visual representations. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. Pmlr. 23, 35

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. E. (2020c). Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255. 11, 12, 23, 35

Chen, X., Fan, H., Girshick, R., and He, K. (2020d). Improved baselines with momentum contrastive learning. In *https://arxiv.org/abs/2003.04297*. 11, 12

Chen, Y., Liu, Y., Jiang, D., Zhang, X., Dai, W., Xiong, H., and Tian, Q. (2022). Sdae: Self-distillated masked autoencoder. In *European Conference on Computer Vision*, pages 108–124. Springer. 53

Chen, Z., Badrinarayanan, V., Lee, C.-Y., and Rabinovich, A. (2018). Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pages 794–803. PMLR. 49

Cheng, G., Han, J., and Lu, X. (2017a). Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883. 38

Cheng, G., Xie, X., Han, J., Guo, L., and Xia, G.-S. (2017b). Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883. 3

Christie, G., Fendley, N., Wilson, J., and Mukherjee, R. (2018). Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180. 19

Cong, Y., Khanna, S., Meng, C., Liu, P., Rozi, E., He, Y., Burke, M., Lobell, D. B., and Ermon, S. (2022). Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *arXiv preprint arXiv:2207.08051*. 3, 13, 31

Darcet, T., Oquab, M., Mairal, J., and Bojanowski, P. (2023). Vision transformers need registers. *arXiv preprint arXiv:2309.16588*. 4, 16, 57, 60, 69

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. 1, 3

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. 11

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. 15

Do, K. and Tran, T. (2019). Theory and evaluation metrics for learning disentangled representations. *arXiv preprint arXiv:1908.09961*. 10

Do, K. and Tran, T. (2021). Theory and evaluation metrics for learning disentangled representations. 65

Dollár, P. and Zitnick, C. L. (2014). Fast edge detection using structured forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(8):1558–1570. 13

Dong, S., Hangel, G., Bogner, W., Widhalm, G., Rössler, K., Trattnig, S., You, C., de Graaf, R., Onofrey, J. A., and Duncan, J. S. (2022). Multi-scale super-resolution magnetic resonance spectroscopic imaging with adjustable sharpness. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 410–420. Springer. 13

Dong, X., Bao, J., Zheng, Y., Zhang, T., Chen, D., Yang, H., Zeng, M., Zhang, W., Yuan, L., Chen, D., et al. (2023). Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10995–11005. 53

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. 8, 11, 13, 35

Eastwood, C. and Williams, C. K. (2018). A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*. 4

Farabet, C., Couprie, C., Najman, L., and LeCun, Y. (2012). Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1915–1929. 13

Felzenszwalb, P., McAllester, D., and Ramanan, D. (2008). A discriminatively trained, multiscale, deformable part model. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. Ieee. 13

Feng, Z. and Zhang, S. (2023). Evolved part masking for self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10386–10395. 15

Georgiou, K., Tang, M., and Qi, H. (2026). Explore: Expert patch-level loss routing for multi-objective masked image modeling. 3, 5, 46, 48, 74, 79

Georgiou, K., **Tang**, t., Wang, F., Tang, W., Qi, H., Champion, C., and Bosch, M. (2024). Advancing multi-scale remote sensing analysis through self-supervised learning fine-tuning strategies. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 7031–7034. IEEE. 3, 5, 30, 44, 79

Greff, K., Kaufman, R. L., Kabra, R., Watters, N., Burgess, C., Zoran, D., Matthey, L., Botvinick, M., and Lerchner, A. (2020). Multi-object representation learning with iterative variational inference. 17

Grill, J.-B., Strub, F., Altch'e, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. ., Guo, Z., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M. (2020). Bootstrap your own latent: A new approach to self-supervised learning. *Neural Information Processing Systems*. 1, 14

He, J., Yang, S., Yang, S., Kortylewski, A., Yuan, X., Chen, J.-N., Liu, S., Yang, C., Yu, Q., and Yuille, A. (2022a). Partimagenet: A large, high-quality dataset of parts. In *European Conference on Computer Vision*, pages 128–145. Springer. 66

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022b). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009. 2, 3, 9, 11, 14, 20, 30, 41, 47, 53

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020a). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738. 1, 20

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., Xun, H., and Belongie, S. (2020b). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*. 11, 12

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778. 1, 22

Helber, P., Bischke, B., Dengel, A., and Borth, D. (2019). Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226. 38

Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., and Lerchner, A. (2018). Towards a definition of disentangled representations. In *Workshop on Disentanglement at NeurIPS*. 4

Hou, Z., Sun, F., Chen, Y.-K., Xie, Y., and Kung, S.-Y. (2022). Milan: Masked image pretraining on language assisted representation. *arXiv preprint arXiv:2208.06049*. 14, 53

Ji, P., Yan, S., and Liu, Q. (2013). Region-based spatial sampling for image classification. In *2013 Seventh International Conference on Image and Graphics*, pages 874–879. IEEE. 12

Jia, M., Tang, L., Chen, B.-C., Cardie, C., Belongie, S., Hariharan, B., and Lim, S.-N. (2022). Visual prompt tuning. 15

Jiao, L., Gao, J., Liu, X., Liu, F., Yang, S., and Hou, B. (2021). Multi-scale representation learning for image classification: A survey. *IEEE Transactions on Artificial Intelligence*. 13

Jing, L. and Tian, Y. (2020). Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4037–4058. 1

Kakogeorgiou, I., Gidaris, S., Psomas, B., Avrithis, Y., Bursuc, A., Karantzalos, K., and Komodakis, N. (2022). What to hide from your students: Attention-guided masked image modeling. In *European Conference on Computer Vision*, pages 300–318. Springer. 14

Kim, H. and Mnih, A. (2018). Disentangling by factorising. In *International conference on machine learning*, pages 2649–2658. PMLR. 57

Kovashka, A. and Grauman, K. (2010). Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *2010 IEEE International Conference on Computer Vision and Pattern Recognition*, pages 2046–2053. IEEE. 13

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105. 1

Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006*

IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'06), volume 2, pages 2169–2178. IEEE. 12

Li, G., Zheng, H., Liu, D., Wang, C., Su, B., and Zheng, C. (2022a). Semmae: Semantic-guided masking for learning masked autoencoders. *Advances in Neural Information Processing Systems*, 35:14290–14302. 14

Li, H., Li, Y., Zhang, G., Liu, R., Huang, H., Zhu, Q., and Tao, C. (2022b). Global and local contrastive self-supervised learning for semantic segmentation of hr remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, page 1–1. 12

Lin, G., Shen, C., Van Den Hengel, A., and Reid, I. (2016). Efficient piecewise training of deep structured models for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3194–3203. 13

Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., and Tang, J. (2021). Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1):857–876. 1

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. (2019). Challenging common assumptions in the unsupervised learning of disentangled representations. In *International conference on machine learning*, pages 4114–4124. 4

Long, S., Zhao, Z., Pi, J., Wang, S., and Wang, J. (2022). Beyond attentive tokens: Incorporating token importance and diversity for efficient vision transformers. 15

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110. 1

Manas, O., Lacoste, A., Giró-i Nieto, X., Vazquez, D., and Rodriguez, P. (2021). Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data.

In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9414–9423. 11

Mansouri, A., Hartford, J., Zhang, Y., and Bengio, Y. (2023). Object-centric architectures enable efficient causal representation learning. 17

Mendieta, M., Han, B., Shi, X., Zhu, Y., Chen, C., and Li, M. (2023). Gfm: Building geospatial foundation models via continual pretraining. *arXiv preprint arXiv:2302.04476*. 11

Miao, L. and Qi, H. (2007). Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization. *IEEE Transactions on Geoscience and Remote Sensing*, 45(3):765–777. 61

Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*. 35

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. (2024). Dinov2: Learning robust visual features without supervision. 16

Peebles, W. and Xie, S. (2023). Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205. 75

Peng, Z., Dong, L., Bao, H., Ye, Q., and Wei, F. (2022). A unified view of masked image modeling. *arXiv preprint arXiv:2210.10615*. 53

Pohlen, T., Hermans, A., Mathias, M., and Leibe, B. (2017). Full-resolution residual networks for semantic segmentation in street scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4151–4160. 13

Puigcerver, J., Riquelme Ruiz, C., Mustafa, B., and Houlsby, N. (2024). From sparse to soft mixtures of experts. In *International Conference on Learning Representations*, volume 2024, pages 28435–28445. 49, 76

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*. 2, 47, 48

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training. *OpenAI Blog*. 11

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9. 11

Reed, C. J., Gupta, R., Li, S., Brockman, S., Funk, C., Clipp, B., Candido, S., Uyttendaele, M., and Darrell, T. (2022). Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. *arXiv preprint arXiv:2212.14532*. 3, 13, 31, 39

Reininghaus, J., Huber, S., Bauer, U., and Kwitt, R. (2015). A stable multi-scale kernel for topological machine learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4741–4748. 13

Ren, S., Zhou, D., He, S., Feng, J., and Wang, X. (2022). Shunted self-attention via multi-scale token aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10853–10862. 13

Rosenfeld, A. and Thurston, M. (1971). Edge and curve detection for visual scene analysis. *IEEE Transactions on Computers*, 100(5):562–569. 12

Seitzer, M., Horn, M., Zadaianchuk, A., Zietlow, D., Xiao, T., Simon-Gabriel, C.-J., He, T., Zhang, Z., Schölkopf, B., Brox, T., et al. (2023). Bridging the gap

to real-world object-centric learning. In *International Conference on Learning Representations*. 4, 57

Shi, Y., Siddharth, N., Torr, P., and Kosiorek, A. R. (2022). Adversarial masking for self-supervised learning. In *International Conference on Machine Learning*, pages 20026–20040. PMLR. 14

Singh, B. and Davis, L. S. (2018). An analysis of scale invariance in object detection snip. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3578–3587. 13

Singh, G., Wu, Y.-F., and Ahn, S. (2022). Simple unsupervised object-centric learning for complex and naturalistic videos. 17

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9. 13

Tang, M., Cozma, A., Georgiou, K., and Qi, H. (2023a). Cross-scale mae: A tale of multiscale exploitation in remote sensing. *Advances in Neural Information Processing Systems*, 36:20054–20066. 3, 5, 30, 39, 73, 78

Tang, M., Georgiou, K., Qi, H., Champion, C., and Bosch, M. (2023b). Semantic segmentation in aerial imagery using multi-level contrastive learning with local consistency. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3798–3807. 3, 5, 11, 20, 73, 78

Tang, M. and Qi, H. (2026). The extra tokens matter: Disentangled representation learning with vision transformers. In *Forty-third International Conference on Machine Learning*. 4, 5, 17, 56, 66, 68, 74, 79

Tian, Y., Krishnan, D., and Isola, P. (2020). Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer. 35

Wang, X., Zhang, R., Shen, C., Kong, T., and Li, L. (2021). Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3024–3033. 21

Wang, Y., Cheng, L., Fang, C., Zhang, D., Duan, M., and Wang, M. (2024a). Revisiting the power of prompt for visual tuning. 15

Wang, Y., Zhang, Q., Guo, Y., and Wang, Y. (2024b). Non-negative contrastive learning. 65

Wei, C., Fan, H., Xie, S., Wu, C.-Y., Yuille, A., and Feichtenhofer, C. (2022). Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678. 2, 47

Woo, J., Stone, M., and Prince, J. L. (2015). Multimodal registration via mutual information incorporating geometric and spatial context. *IEEE Transactions on Image Processing*, 24(2):757–769. 26

Xia, G.-S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., and Lu, X. (2017). Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981. 38

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., and Hu, H. (2022). Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663. 14, 47

Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., and Wang, X. (2022). Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18134–18144. 4, 57

Yan, S., Xu, X., Xu, D., Lin, S., and Li, X. (2012). Beyond spatial pyramids: A new feature extraction framework with dense spatial sampling for image classification. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV 12*, pages 473–487. Springer. 12

Yan, S., Xu, X., Xu, D., Lin, S., and Li, X. (2014). Image classification with densely sampled image windows and generalized adaptive multiple kernel learning. *IEEE Transactions on Cybernetics*, 45(3):381–390. 12

Yin, H., Vahdat, A., Alvarez, J., Mallya, A., Kautz, J., and Molchanov, P. (2022). Adavit: Adaptive tokens for efficient vision transformer. 15

Yoo, S., Kim, E., Jung, D., Lee, J., and Yoon, S. (2023). Improving visual prompt tuning for self-supervised vision transformers. 15

Yuan, X., Shi, J., and Gu, L. (2021). A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Systems with Applications*, 169:114417. 2

Zhong, X., Gu, C., Ye, M., Huang, W., and Lin, C.-W. (2022). Graph complemented latent representation for few-shot image classification. *IEEE Transactions on Multimedia*. 22

Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., and Kong, T. (2021). ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*. 14

Zitová, B. and Flusser, J. (2003). Image registration methods: a survey. *Image and Vision Computing*, 21(11):977–1000. 26

Zou, Y. and Cheng, L. (2021). A transfer learning model for gesture recognition based on the deep features extracted by cnn. *IEEE Transactions on Artificial Intelligence*, 2(5):447–458. 13

Publications

Publications Related to This Dissertation

- M. Tang and H. Qi, "The extra tokens matter: Disentangled representation learning with vision transformers," in *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026, Accepted.

- K. Georgiou, M. Tang, and H. Qi, "ExPLoRe: Expert patch-level loss routing for multi-objective masked image modeling," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026, Accepted.

- M. Tang, A. Cozma, K. Georgiou, and H. Qi, "Cross-Scale MAE: A tale of multiscale exploitation in remote sensing," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023.

- K. Georgiou, M. Tang, F. Wang, W. Tang, H. Qi, C. Champion, and M. Bosch, "Advancing multi-scale remote sensing analysis through self-supervised learning fine-tuning strategies," in *IGARSS 2024 – IEEE International Geoscience and Remote Sensing Symposium*, 2024, pp. 7031–7034.

- M. Tang, K. Georgiou, H. Qi, C. Champion, and M. Bosch, "Semantic segmentation in aerial imagery using multi-level contrastive learning with local consistency," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023, pp. 3798–3807.

Other Publications

- M. Tang and H. Qi, "Looking into a pixel by nonlinear unmixing – a generative approach," arXiv:2604.01141, 2026.
- K. Georgiou, M. Tang, and H. Qi, "MEDiC: Multi-objective exploration of distillation from CLIP," arXiv:2603.29009, 2025.
- F. Wang, W. Tang, M. Tang, et al., "Koopman-based transition detection in satellite imagery: Unveiling construction phase dynamics through material histogram analysis," in *IGARSS 2024 – IEEE International Geoscience and Remote Sensing Symposium*, 2024, pp. 8786–8790.
- M. Shen, M. Tang, W. Jiao, and Y. Li, "Kudzu invasion and its influential factors in the southeastern United States," *International Journal of Applied Earth Observation and Geoinformation*, vol. 130, 2024, p. 103872.
- M. Shen, M. Tang, and Y. Li, "Phenology and spectral unmixing-based invasive kudzu mapping: A case study in Knox County, Tennessee," *Remote Sensing*, vol. 13, 2021, p. 4551.
- M. Shen, S. Wang, Y. Li, M. Tang, and Y. Ma, "Pattern of turbidity change in the middle reaches of the Yarlung Zangbo River, southern Tibetan Plateau, from 2007 to 2017," *Remote Sensing*, vol. 13, 2021, p. 182.
- M. Tang, Y. Qu, and H. Qi, "Hyperspectral nonlinear unmixing via generative adversarial network," in *IGARSS 2020 – IEEE International Geoscience and Remote Sensing Symposium*, 2020, pp. 2404–2407.
- L. Gao, X. Sun, W. Luo, M. Tang, and B. Zhang, "Swarm intelligence algorithms for spectral unmixing in hyperspectral image," *Journal of Nanjing University of Information Science and Technology*, vol. 10, 2018, pp. 81–91.

- M. Tang, B. Zhang, A. Marinoni, L. Gao, and P. Gamba, "Multiharmonic postnonlinear mixing model for hyperspectral nonlinear unmixing," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, 2018, pp. 1765–1769.

- M. Tang, L. Gao, A. Marinoni, P. Gamba, and B. Zhang, "Integrating spatial information in the normalized p-linear algorithm for nonlinear hyperspectral unmixing," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, 2017, pp. 1179–1190.

- M. Tang, L. Gao, A. Marinoni, and B. Zhang, "Nonlinear hyperspectral unmixing based on normalized p-linear algorithm," in *IGARSS 2017 – IEEE International Geoscience and Remote Sensing Symposium*, 2017, pp. 1145–1148.

Manuscripts in Preparation or Under Review

- M. Tang and H. Qi, "Beyond EMA: Dynamics-aware teacher updates via Koopman operator learning," submitted to the *International Conference on Learning Representations (ICLR)*, 2027.

- M. Tang, K. Georgiou, W. Tang, et al., "Koopman-based transient event detection in temporal satellite imagery," submitted to *IEEE Transactions on Geoscience and Remote Sensing*, 2026.

- M. Tang, Z. Zhang, and H. Qi, "Understanding and stabilizing joint VAE-DiT training: A mechanistic analysis of latent conflicts," submitted to the *Conference on Neural Information Processing Systems (NeurIPS)*, 2026.

- W. Tang, F. Wang, M. Tang, et al., "Enhanced phase transition detection in satellite imagery time series via a Koopman-based framework," submitted to *Nature Communications*, 2026.

Vita

Maofeng Tang received the B.S. degree in Geodesy and Geomatics from Wuhan University, China, in 2015, and the M.S. degree in Remote Sensing from the University of Chinese Academy of Sciences in 2018, where he conducted research on spectral unmixing and spatial analysis in remote sensing under the direction of Dr. Bing Zhanga. He joined the University of Tennessee, Knoxville, in 2018 to pursue the Ph.D. degree in Computer Engineering under the direction of Dr. Hairong Qi.

From 2018 to 2021, he served as a Graduate Teaching Assistant, supporting the Machine Learning, Pattern Recognition, and Deep Learning courses. Since 2021, he has served as a Graduate Research Assistant, leading research efforts on several federally and industry-funded projects, including the IARPA-funded MUTATED program in collaboration with Accenture Federal Services and a research gift from Adobe. He received the Tennessee Fellowship for Graduate Excellence Award in 2022 and 2023 and was selected for the Adobe Research University Gift Program in 2025.

His dissertation research, conducted under the direction of Dr. Hairong Qi, addresses self-supervised visual representation learning through a three-phase research program: domain-specific foundations in remote sensing imagery, generalization of the underlying design principle to natural images through adaptive, patch-level loss routing, and refinement toward representations that are simultaneously interpretable and performant.