

A CONCEPTUAL FRAMEWORK FOR ROBUST FEW-SHOT LEARNING: INTEGRATING UNBALANCED OPTIMAL TRANSPORT AND SELF-SUPERVISED TRANSFORMER REPRESENTATIONS

Hayati Abd Rahman^{1*} and Pang Yun²

^{1*}Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM),
Shah Alam, Malaysia

^{2*}Guilin University of Electronic Technology, Beihai, Guangxi, China

^{1*}hayatiar@tmsk.uitm.edu.my, ²yunpang.guet.edu@gmail.com

(* indicates the corresponding author)

ABSTRACT

Few-shot learning (FSL) aims to enable deep models to generalise from extremely limited labelled data, yet unstable metric matching, distribution imbalances, and weak structural representations in low-data regimes often constrain its performance. This paper proposes a conceptual framework that unifies metric-based similarity learning, Unbalanced Optimal Transport (UOT) via Unbalanced Sinkhorn Distance (USD), and self-supervised Transformer representations to conceptually address the theoretical and structural limitations of existing FSL approaches. The framework theoretically unifies distribution-aware USD matching, SSL-enhanced ViT/Swin feature representations, and metric-based inference within a coherent pipeline. This work aims to provide a theoretical foundation and research roadmap for future empirical studies on robust few-shot learning under realistic, distributionally complex conditions.

Keywords: *Few-Shot Learning, Metric Learning, Optimal Transport, Self-Supervised Learning, Sinkhorn Distance, Unbalanced Vision Transformer.*

Received for review: 06-01-2026; Accepted: 25-03-2026; Published: 01-04-2026

DOI: 10.24191/mjoc.vol11i1.9616

1. Introduction

Few-shot learning (FSL) has emerged as a crucial research paradigm in modern machine learning, particularly in scenarios where label data is scarce, costly to obtain, or inherently imbalanced (Song et al., 2022). From medical imaging and industrial defect detection to personalised recommendations and security surveillance, many real-world environments do not permit the large-scale annotation required by conventional deep learning methods (Alsaleh et al., 2024). Despite remarkable progress in neural architectures and optimisation strategies, most state-of-the-art models still depend heavily on large, annotated datasets to achieve robust generalisation. This inherent dependence underscores the importance of developing learning systems capable of adapting to only a handful of labelled examples that align more closely with human learning behaviour (Hospedales et al., 2022).



This is an open access article under the CC BY-SA license
(<https://creativecommons.org/licenses/by-sa/3.0/>).

Existing approaches to FSL can be broadly categorised as metric-based learning, meta-learning, and fine-tuning strategies (Chamarthi et al., 2024). Among these, metric-based methods such as prototypical networks, matching networks, and relation networks remain a foundational pillar due to their conceptual simplicity and computational efficiency (Wang et al., 2020). These models aim to learn an embedding space in which intra-class similarity is maximised and inter-class similarity is minimised, enabling classification via distance-based matching. However, despite their effectiveness in controlled experimental conditions, metric-based models lack robustness in real-world settings, marked by distribution imbalances, noisy features, and structural inconsistencies between support and query samples (Zhang et al., 2023). This constraint aligns with extensive research in machine learning, indicating that imbalanced or skewed data distributions can destabilize similarity estimation and impair classification accuracy, even in traditional supervised contexts (Beh, et al, 2014). Euclidean and cosine similarity are two examples of classical distance functions that depend on comparing points to each other. They also assume that class distributions are geometrically compact and well-balanced. This assumption rarely holds in practical applications, leading to unstable similarity estimates and mismatches in feature alignment.

To overcome these limitations, optimal transport (OT) has been introduced as a principled mathematical framework for distribution-level matching. OT enables the modelling of set-to-set relationships by computing the minimum transport cost between two distributions, thereby capturing global structural properties that conventional metrics overlook. Methods such as Earth Mover's Distance (EMD) and the Sinkhorn Distance have gained significant traction in few-shot tasks due to their ability to model class structure, handle distribution shifts, and incorporate semantic coupling across samples. However, classical OT assumes strict mass conservation; for instance, the total mass of the source and target distributions must be equal. This assumption is highly restrictive in FSL, where support and query sets often display asymmetric distribution mass, background clutter, or irrelevant feature regions. Enforcing balanced transport in such scenarios results in forced or incorrect matching, ultimately degrading classification performance. While the entropic Sinkhorn variant improves computational tractability, it still inherits mass-conservation rigidity and struggles under imbalanced or noisy distributions (Séjourné et al., 2023).

Unbalanced optimal transport (UOT) has become a strong substitute to address these problems. By applying divergence-based penalisation to marginal deviations, UOT loosens the mass conservation constraint. A notable realisation of UOT is the Unbalanced Sinkhorn Distance (USD), which incorporates Kullback–Leibler divergence terms to allow selective matching and controlled mass variation. USD lets the transport plan ignore areas of low quality, give less weight to features that aren't important, and align distributions in a more flexible way. This makes it especially useful for few-shot situations where there are a structural mismatch and not enough data. Despite its theoretical advantages, USD has yet to be systematically integrated into a broader conceptual framework for FSL, especially one that also addresses representation-level challenges.

Parallel to the development of OT-based methods, self-supervised learning (SSL) has revolutionised representation learning in vision systems by enabling feature extraction from large amounts of unlabelled data. Teacher–student SSL paradigms such as BYOL, DINO, and iBOT have demonstrated strong capability in producing semantically rich and highly transferable representations, outperforming supervised pretraining in several low-shot and cross-domain settings. With the advent of Vision Transformers (ViT) and Swin Transformers, SSL methods have become even more powerful, providing global and local contextual modelling that can greatly enhance few-shot generalisation. However, transformers are highly data-dependent and structurally sensitive under low-data regimes, limiting their adaptability when they are not properly regularised or enhanced. Although SSL offers a strong foundation for generating high-quality visual representations, its integration with distribution-aware metric learning, particularly UOT, remains underexplored.

Despite notable advancements in metric-based classifiers, optimal transport formulations, and self-supervised Transformer architectures, the existing literature provides no comprehensive theoretical framework that integrates these developments into a unified perspective. While unbalanced optimal transport techniques such as UOT and USD offer principled, distribution-aware matching that is resilient to imbalance and noise, and self-supervised Transformers yield semantically rich and structurally consistent representations suitable for low-data environments, these contributions remain conceptually isolated. Furthermore, none of these approaches has been systematically aligned with the metric-based inference mechanisms that underpin the N-way K-shot paradigm. The absence of an integrative model that concurrently addresses distributional alignment, structural representation learning, and episodic decision-making constitutes a critical theoretical gap. This fragmentation underscores the necessity for a coherent conceptual framework capable of synthesizing these complementary strands into a unified and theoretically grounded pipeline for robust few-shot learning.

Accordingly, the main conceptual contributions of this paper are as follows:

- First, this paper proposes a unified conceptual framework that integrates Unbalanced Sinkhorn Distance (USD) with self-supervised transformer-based representations to enable structurally adaptive and distribution-aware few-shot learning. The framework aims to address the instability of conventional metric learning under imbalanced feature distributions by incorporating flexible transport-based similarity modelling.
- Second, this study provides a theoretical integration of metric learning, optimal transport theory, and teacher–student self-supervised learning paradigms, demonstrating how these approaches can complement each other in low-data regimes. This synthesis highlights the role of distribution-level alignment and representation learning in improving the robustness of few-shot classification.
- Third, the paper systematically analyses limitations in the existing few-shot learning literature and identifies key research gaps related to distribution imbalance, feature misalignment, and the lack of unified learning pipelines. Based on this analysis, the proposed framework provides a conceptual foundation for future empirical and methodological developments in robust few-shot learning.

This study is conceptual and theoretical in nature and does not present experimental validation. Instead, it synthesises existing advances in metric learning, unbalanced optimal transport, and self-supervised Transformer representations into a unified framework intended to serve as a foundation for future empirical research and model development in few-shot learning.

2. Literature Review

Few-shot learning (FSL) has emerged as a crucial paradigm as modern applications increasingly face constraints in labelled data availability. Traditional deep learning models depend on abundant annotated samples, leading to degraded performance when training data is limited or imbalanced. FSL overcomes the problem by adopting N-way K-shot episodic tasks, enabling generalisation from minimal supervision (Wang et al., 2020). The episodic training strategy simulates low-data test conditions, improving cross-task transferability (Hospedales et al., 2022). This section reviews three major theoretical pillars underpinning modern FSL: metric learning, optimal transport, and self-supervised Transformer representations.

2.1 Metric Learning in Few-Shot Classification

Metric learning is one of the first and most important parts of FSL. Models such as Matching Networks (Vinyals et al., 2016), Prototypical Networks (Snell et al., 2017), and Relation Networks (Sung et al., 2018) demonstrate that classification can be achieved through comparing embedded support and query samples via distance functions. These approaches assume that intraclass samples form compact clusters and that simple metrics, like Euclidean or cosine, represent class semantics (Wang et al., 2020).

However, real-world FSL faces distribution mismatch, background clutter, and high intra-class variability, making point-level metrics unstable (Chen et al., 2019). Embedding-based methods like Meta-Baseline (Chen et al., 2021), DeepEMD (Sung et al., 2018) and set-to-set adaptation models (Ye et al., 2020) highlight that simple geometric distances often fail under fine-grained variation. Studies further demonstrate that forced metric matching under imbalance produces unreliable similarity scores (Li et al., 2020). These limitations motivate the shift toward *distribution-level* similarity functions rather than point-to-point comparisons.

2.2 Optimal Transport for Distribution-Level Similarity

Optimal Transport (OT) provides a mathematically principled framework for measuring differences between distributions. Unlike point-based metrics, OT captures global structure by computing the minimal cost to transform one distribution into another (Feydy et al., 2019).

2.2.1 Classical OT

Classical OT formulates alignment as a linear programming problem, offering expressive transport plans but at a cubic computational cost (Feydy, et. al. 2019). Its strict mass-conservation constraint forces support and query features to match exactly, even when distributions are noisy or asymmetric, causing misalignment and overfitting (Chizat, et. al., 2018).

2.2.2 Balanced Sinkhorn Distance

Sinkhorn Distance incorporates entropic regularization to obtain a computationally tractable approximation of optimal transport, enabling stable and efficient iterative updates on modern GPU architectures. Nevertheless, the formulation retains the fundamental mass-conservation constraint inherent to classical optimal transport, which results in sensitivity to noise, background clutter, and disparities in support–query distribution densities (Feydy et al., 2019).

2.2.3 Unbalanced Optimal Transport (UOT) and Unbalanced Sinkhorn Distance (USD)

UOT relaxes mass conservation using divergence penalties, which are typically KL divergence, that allow partial matching and flexible mass variation (Chizat et al., 2018). USD applies these principles to Sinkhorn iterations, enabling:

- selective matching of semantic regions,
- down-weighting of irrelevant features,
- robust handling of imbalanced distributions, and
- improved stability under noise or cross-domain shifts (Chizat et al., 2018).

Despite these strengths, USD remains underexplored in FSL pipelines, particularly in combination with modern representation learning approaches.

2.3 Self-Supervised Learning for Few-Shot Representation

Self-supervised learning (SSL) has transformed visual representation learning by enabling models to extract semantics directly from unlabelled data. SSL is especially relevant in FSL, where label scarcity often limits generalisation and introduces overfitting (Wang et al., 2020; Gao et al., 2021).

Teacher–student frameworks such as BYOL, DINO, and iBOT learn discriminative features through multi-view consistency and patch-level alignment. DINO, which is based on ViT, creates representation spaces that are well-clustered and good for downstream FSL tasks (Caron et al., 2021). iBOT extends this approach by enforcing instance- and patch-level uniformity, enhancing structural consistency under low-data conditions (Zhou et al. 2022). Meta-Self further validates the capability of SSL features to improve few-shot generalisation (Yoon et al., 2025).

2.3.1 Vision Transformer (ViT)

ViT tokenises images into patches and applies global self-attention, enabling strong global semantic modelling (Dosovitskiy et al., 2021). However, ViT is data-hungry and often unstable during early training when labelled samples are scarce (Caron et al., 2021; Liu et al., 2021).

2.3.2 Swin Transformer

Swin incorporates window-based self-attention (W-MSA), improving local structural modelling but limiting global context due to restricted inter-window information flow (Dosovitskiy et al., 2021).

2.3.3 SSL-Enhanced Transformer Representation

Self-supervised learning (SSL) has been shown to alleviate several structural limitations inherent in Transformer-based architectures by promoting consistent alignment between global and local features across augmented views, enforcing patch-level structural coherence, and improving class separability in low-shot environments (Huang et al., 2022). These enhancements contribute to more stable and discriminative representations under limited supervision. However, despite these developments, SSL-driven Transformer models have not yet been systematically integrated with distribution-aware similarity measures such as the Unbalanced Sinkhorn Distance (USD), leaving a gap in the formulation of unified frameworks for robust few-shot learning.

2.4 Findings from Prior Work

Although metric learning, OT-based approaches, and self-supervised transformers have each contributed meaningful progress to few-shot learning, the existing literature remains conceptually fragmented. The first limitation concerns the instability of point-based metric functions under real-world distribution imbalance. When support and query samples differ in density, contain clutter, or exhibit asymmetric semantic structure, classical distances degrade significantly, resulting in unreliable similarity estimation (Li et al., 2020). The second limitation is the weakness that arises from structural constraints within transformer-based representations. Vision Transformer (ViT) models struggle to capture fine-grained local cues, while Swin Transformer sacrifices global contextual breadth; although SSL techniques such as DINO and iBOT partially alleviate these issues, they do not fully resolve structural inconsistencies between support and query embeddings (Caron et al. 2021; Zhou et al. 2022).

The third limitation is the absence of a unified framework that integrates SSL-enhanced Transformer features, explicit patch-level structural alignment, and distribution-aware matching via Unbalanced Optimal Transport. These components have evolved in parallel rather

than in concert, leaving a theoretical disconnect in how modern FSL models should jointly learn representations and match distributions. This fragmentation underscores the need for the unified conceptual framework proposed in Section 4, which consolidates SSL Transformers, structural alignment, and USD-driven similarity estimation to address the core representational and distributional weaknesses inherent in existing FSL approaches.

2.5 Research Gaps and Motivation

Although metric learning, optimal transport, and self-supervised Transformers have significantly strengthened the foundations of few-shot learning, the literature demonstrates several unresolved limitations. Through a synthesis of prior work in Section 2, three major gaps emerge.

2.5.1 Instability of Point-Based Metric Learning Under Distribution Imbalance

Classical metric-based methods assume compact intra-class clustering and structural consistency between support and query representations. However, empirical findings show that Euclidean and cosine metrics degrade under distributional asymmetry, noisy regions, cluttered backgrounds, and fine-grained intra-class variations (Wang et al., 2020; Zhang et al., 2023; Li et al., 2020). Even advanced embedding-based methods such as Meta-Baseline (Chen et al., 2021) and DeepEMD (Sung et al., 2018) struggle to ensure stable matching when semantic density varies across samples.

Existing work attempts to address such challenges through set-to-set matching (Ye et al., 2020) or attention-based alignment (Huang et al., 2022), but these solutions do not fundamentally solve the problem of forced matching caused by mass-conserving distance functions. This highlights the need for distribution-aware similarity measures capable of selective alignment—an ability provided by unbalanced OT formulations.

2.5.2 Structural Limitations of Transformer Representations in Low-Data Regimes

Despite their success in large-scale vision tasks, Transformer architectures face structural difficulties in few-shot settings. ViT is known to require substantial training data to stabilise global attention maps (Dosovitskiy et al., 2021; Liu et al., 2021; Caron et al., 2021), while Swin Transformer sacrifices global receptive field due to window-based partitioning, limiting long-range semantic reasoning (Dosovitskiy et al., 2021).

SSL partially reduces these limitations through multi-view consistency and patch-level reconstruction. Methods such as DINO (Caron et al., 2021) and iBOT (Zhou et al., 2022) enhance representation coherence but do not explicitly address geometric misalignment between support and query samples. Moreover, these models operate primarily at the feature extraction stage and do not incorporate distribution-level matching, leaving misalignment to downstream similarity functions that remain point-based and unstable.

2.5.3 Lack of a Unified Framework Integrating SSL, Structural Alignment, and Unbalanced Optimal Transport

Present literature treats SSL representation, structural patch alignment, and distribution-aware similarity as separate advancements. No existing framework integrates:

- SSL-driven high-level semantics (e.g., DINO, iBOT),
- explicit structural alignment modules, and
- flexible distribution matching via unbalanced Sinkhorn distance (Chizat et al., 2018).

While UOT/USD has been shown to outperform classical OT under imbalanced distributions (Chizat et al., 2018), it has never been combined with SSL-Transformer features

nor incorporated into a unified FSL architecture. This disconnect represents a major theoretical gap that limits further progress. Thus, a conceptual framework that unifies representation, structure, and distribution matching is urgently needed—and this aspect motivates the proposed model in Section 3.

Table 1. Summary of Literature Review and Research Gap in Few-Shot Learning

Research Direction	Representative Work	Limitation	Research Gap
Metric-based learning	Matching Networks, ProtoNet, RelationNet	Sensitive to distribution imbalance and noisy features	Need distribution-aware similarity modelling
Optimal Transport	Sinkhorn Distance, DeepEMD	Strict mass conservation leads to forced matching	Need flexible transport under imbalanced distributions
Self-Supervised Transformers	DINO, iBOT	Strong representations but lack explicit similarity modelling	Need integration with distribution-aware matching
This Study	SSL Transformer + USD	-	Unified framework integrating representation learning and distribution-aware similarity

Table 1 summarises the main research directions in few-shot learning and highlights their key limitations. Metric-based methods rely on point-level similarity and are sensitive to distribution imbalance. Optimal transport improves distribution-level matching but typically assumes strict mass conservation. Meanwhile, self-supervised Transformer models provide powerful representations but lack explicit mechanisms for distribution-aware similarity modelling. These limitations reveal a clear research gap: the absence of a unified framework that integrates robust representation learning with flexible distribution matching. The conceptual framework proposed in this study, described in the following section, aims to address this gap.

3. Proposed Conceptual Framework

The limitations identified in the preceding sections motivate the development of a unified conceptual framework that combines self-supervised Transformer representations, structural alignment mechanisms, and distribution-aware matching through the Unbalanced Sinkhorn Distance (USD). This integrated design aims to stabilise feature learning under low-data conditions, strengthen semantic coherence between support and query samples, and improve robustness in similarity estimation through flexible, mass-relaxed transport. The overall architecture of the proposed framework is presented in Figure 1.

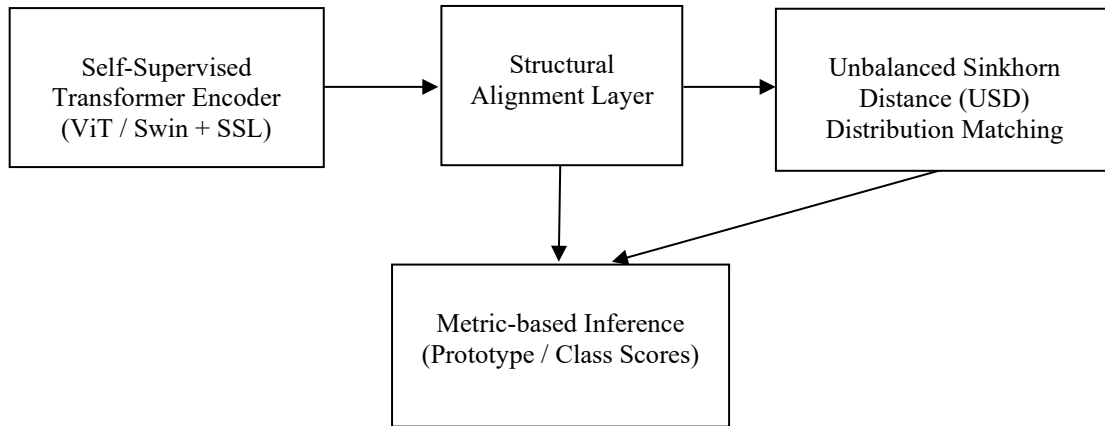


Figure 1. Conceptual Framework for Robust Few-Shot Learning (SSL Transformer + USD).

3.1 Self-Supervised Transformer Encoder (ViT/Swin + SSL)

The first component employs Vision Transformer (ViT) (Dosovitskiy et al., 2021) or Swin Transformer as the backbone, pretrained under a self-supervised paradigm such as DINO (Caron et al., 2021) or iBOT (Zhou et al., 2022). SSL produces representations that are semantically enriched, invariant across augmentations, and robust to label scarcity.

DINO provides globally coherent attention maps suitable for few-shot generalisation, while iBOT additionally enforces patch-level alignment, improving discrimination in fine-grained tasks. These properties make SSL-based Transformers ideal for serving as the feature extractor in the proposed framework.

3.2 Structural Alignment Layer

Despite strong representation learning, ViT and Swin encounter structural inconsistencies under low-shot conditions (Dosovitskiy et al., 2021; Liu et al., 2021). The structural alignment layer aims to mitigate these issues by normalising patch embeddings, enhancing local structural cues, and preserving global semantics while stabilising spatial patterns.

Methods such as local attention refinement (Yoon et al., 2025) and patch-level contrastive alignment (Zhou et al., 2022) inspire the design of this alignment stage. This layer serves as an intermediate bridge connecting SSL features with distribution-level similarity estimation.

3.3 Unbalanced Sinkhorn Distance for Distribution-Aware Matching

The core component of the framework is the Unbalanced Sinkhorn Distance module, grounded in unbalanced optimal transport theory (Chizat et al., 2018). USD introduces KL divergence penalties to relax mass conservation, enabling flexible alignment that down-weights irrelevant or noisy regions, aligns semantically meaningful areas selectively, avoids forced matching between distributions of unequal mass, and improves robustness under cross-domain and low-data shifts.

Unlike balanced OT or point-based metrics, USD can model partial transport as a property essential for few-shot support-query matching, where structural asymmetry is common (Chizat et al., 2018).

3.4 Metric-Based Inference Layer

After USD produces a distribution-aware similarity score or aligned representation map, classification proceeds through standard metric-based inference. For example,

- Prototypes may be computed from the aligned support embeddings,
- Class scores may be derived directly from USD-based distances.

This final stage aligns with widely adopted episodic FSL procedures (Wang et al., 2020; Snell et al., 2016), ensuring compatibility with existing benchmarks while benefiting from the enhanced robustness introduced by the SSL + USD integration.

4. Implications

The proposed conceptual framework offers several important implications for theory, methodology, and practical deployment of few-shot learning systems. By integrating self-supervised Transformer representations, structural alignment mechanisms, and Unbalanced Sinkhorn Distance (USD)-based distribution matching, the framework provides a structured lens for advancing the next generation of FSL research.

4.1 Theoretical Implications

First, the framework adds a unified theoretical foundation that bridges three previously disconnected research domains: representation learning, structural alignment, and distribution-level similarity modelling. While prior work has independently improved FSL through enhanced embeddings or more expressive similarity functions, no existing theory integrates these components into a single coherent pipeline. The introduction of USD as a core similarity function advances the theoretical understanding of how flexible mass transport and partial matching can mitigate the weaknesses of point-level metrics.

Second, the incorporation of self-supervised Transformers highlights a new direction for representation learning in low-data regimes. SSL shifts reliance away from supervised labels and encourages models to learn intrinsic semantic structures, thereby strengthening generalisation under a few-shot constraint.

Finally, the alignment layer introduces a theoretical connection between patch-level coherence and distribution-aware semantics, suggesting that the optimal FSL model must jointly reason about where information is located (structure) and how it compares across samples (distribution).

4.2 Practical Implications

From an applied perspective, the conceptual framework has strong potential across domains that naturally encounter label shortages, such as medical imaging, industrial inspection, remote sensing, and personalised recommendation systems. In these fields, structural noise, domain shifts, and highly imbalanced distributions are common.

The combination of SSL and USD offers robustness to noise, a better generalisation across domains, strong performance in imbalanced or fine-grained settings, and a scalable pathway for real-world deployment. Practitioners can adopt this framework as a guide for selecting architectures and similarity measures that match operational constraints.

4.3 Research Opportunities

This conceptual framework also opens several promising avenues for future research. One key direction involves exploring joint SSL–UOT training mechanisms, where representation learning and distribution matching are co-optimised to produce more coherent and semantically aligned embeddings. Another opportunity lies in developing lightweight approximations of the Unbalanced Sinkhorn Distance (USD) to enhance computational efficiency, enabling its deployment in real-time or edge-based applications. The framework additionally offers potential for cross-domain few-shot learning studies, particularly in medical imaging, industrial inspection, and other fields where label scarcity and distributional shifts are common, making the combination of SSL features and distribution-aware matching especially valuable. Beyond classification, the principles embedded in USD also lend themselves to extensions in few-shot object detection and segmentation, where partial matching and structural alignment are crucial for structured-output tasks. Collectively, these opportunities position the proposed framework as a solid foundation for advancing empirical work in next-generation few-shot learning.

5. Conclusion

This paper presents a unified conceptual framework for robust few-shot learning that integrates three complementary components: self-supervised Transformer representations, structural alignment mechanisms, and Unbalanced Sinkhorn Distance (USD)-based distribution matching. The framework provides a theoretical basis for addressing key limitations, including metric instability under distribution imbalance, structural weaknesses in Transformer backbones when trained with limited data, and the absence of integrated representation–metric pipelines.

By synthesising insights from metric learning, optimal transport theory, and modern self-supervised learning, the proposed model offers a new theoretical perspective on how robust generalisation can be achieved under low data constraints. The framework highlights the importance of combining semantic-rich, SSL-driven features with distribution-aware similarity measures capable of selective mass transport.

Although conceptual in nature, this work establishes a strong foundation for future empirical studies. Researchers may adopt the framework as a guideline for designing hybrid FSL models that are both structurally adaptive and distributionally robust. Practitioners may leverage its insights to build real-world systems that can operate reliably in environments where annotated data is scarce, noisy, or highly imbalanced.

Overall, the proposed conceptual model represents a meaningful step toward unifying representation learning, structural alignment, and optimal transport-based matching in a coherent theoretical perspective for next-generation few-shot classification.

Acknowledgement

The authors gratefully acknowledge the Faculty of Computer and Mathematical Sciences for providing the opportunity to contribute to this study.

Funding

The authors received no specific grant from any funding agency, commercial or not-for-profit sectors, for this research.

Author Contribution

Pang Yun was responsible for conceptualising this study, methodology, software development, and writing the original draft. Hayati Abd Rahman was responsible for writing this paper based on the original draft, data analysis, and review. All authors had approved the final version.

Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.

References

- Alsaleh, A. M., Albalawi, E., Alghosaibi, A., Albakheet, S. S., & Khan, S. B. (2024). Few-Shot Learning for Medical Image Segmentation Using 3D U-Net and Model-Agnostic Meta-Learning (MAML). *Diagnostics*, 14(12), 1213. <https://doi.org/10.3390/diagnostics14121213>
- Beh, T. Y. K., Tan, S. C. & Yeo, H. T. (2014, February). Building Classification Models from Imbalanced Fraud Detection Data. *Malaysian Journal of Computing*, 2(2).
- Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J. & Bojanowski, P. (2021, October). Emerging Properties in Self-Supervised Vision Transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 9630-9640. doi: 10.1109/ICCV48922.2021.00951
- Chamarthi, S., Fogelberg, K., Gawlikowski, J., & Brinker, T. J. (2024). Few-shot learning for skin lesion classification: A prototypical networks approach. *Informatics in Medicine Unlocked*, 48, 101520. <https://doi.org/10.1016/j.imu.2024.101520>
- Chen, W.-Y., Liu, Y.-C., Kira, Z., Wang, Y.-C. F. & Huang, J.-B. (2019). A Closer Look at Few-shot Classification. *International Conference on Learning Representations*. doi: 10.48550/arXiv.1904.04232
- Chen, Y., Liu, Z., Xu, H., Darrell, T. & Wang, X. (2021, August). Meta-Baseline: Exploring Simple Meta-Learning for Few-Shot Learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021 IEEE Conference on* (pp. 9042–9051). *IEEE*. doi: 10.1109/ICCV48922.2021.00893.

- Chizat, L., Peyré, G., Schmitzer, B. & Vialard, F.-X. (2018, February). Scaling Algorithms for Unbalanced Optimal Transport Problems. *Mathematics of Computing*, 87 (314), pp. 2563–2609. doi: 10.1090/mcom/3303.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. & Houlsby, N. (2021, June). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations (ICLR)*. doi: [10.48550/arXiv.2010.11929](https://doi.org/10.48550/arXiv.2010.11929).
- Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S., Trouvé, A. & Peyré, G. (2019). Interpolating between Optimal Transport and MMD using Sinkhorn Divergences. *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019*, 89. doi: 10.48550/arXiv.1810.08278.
- Hospedales, T., Antoniou, A., Micaelli, P. & Storkey, A. (2022, September). Meta-Learning in Neural Networks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9), 5149–5169. doi: 10.1109/TPAMI.2021.3079209
- Huang, G., Laradji, I., Vazquez, D., Lacoste-Julien, S. & Rodriguez, P. (2022, August). A Survey of Self-Supervised and Few-Shot Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4), pp. 4071 – 4089. doi: 10.1109/TPAMI.2022.3199617.
- Li, W., Wang, L., Huo, J., Shi, Y., Gao, Y. & Luo, J. (2020, February). Asymmetric Distribution Measure for Few-shot Learning. *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*. (pp. 2957-2963). doi: 10.48550/arXiv.2002.00153.
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y. & Zhang, Z. (2021, October). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp.9992-10002. doi: 10.1109/ICCV48922.2021.00986.
- Gao, T., Fisch, A. & Chen, D. (2021, August). Making Pre-trained Language Models Better Few-shot Learners. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. pp. 3816–3830. doi: 10.18653/v1/2021.acl-long.295
- Séjourné, T., Feydy, J., Vialard, F.-X., Trouvé, A., & Peyré, G. (2023). *Sinkhorn Divergences for Unbalanced Optimal Transport* (arXiv:1910.12958). arXiv. <https://doi.org/10.48550/arXiv.1910.12958>
- Snell, J., Swersky, K. & Zemel R. (2017, March). Prototypical Networks for Few-shot Learning. *Advances in Neural Information Processing Systems*. doi: 10.48550/arXiv.1703.05175
- Song, Y., Wang, T., Mondal, S. K., & Sahoo, J. P. (2022). *A Comprehensive Survey of Few-shot Learning: Evolution, Applications, Challenges, and Opportunities* (arXiv:2205.06743). arXiv. <https://doi.org/10.48550/arXiv.2205.06743>
- Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H. S. & Hospedales, T. M. (2018, June). Learning to Compare: Relation Network for Few-Shot Learning. In *2018 IEEE/CVF*

Conference on Computer Vision and Pattern Recognition. 2018 IEEE Conference on (pp. 1199–1208). IEEE. doi: 10.1109/CVPR.2018.00131.

Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K. & Wierstra, D. (2016, December). Matching Networks for One Shot Learning. *30th Conference on Neural Information Processing Systems (NIPS 2016)*. (pp. 3637-3645). doi: 10.48550/arXiv.1606.04080.

Wang, Y., Yao, Q., Kwok, J., & Ni, L. M. (2020, March). Generalizing from a Few Examples: A Survey on Few-Shot Learning. *ACM Computing Surveys (CSUR)*, 53(3), 1-34. doi: 10.1145/3386252.

Ye, H.-J., Hu, H., Zhan, D.-C. & Sha, F. (2020, June). Few-Shot Learning via Embedding Adaptation with Set-to-Set Functions. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Conference on (pp. 8805-8814). IEEE. doi: 10.1109/CVPR42600.2020.00883.

Yoon, H., Kwak, J., Tolera, B. A., & Dai, G. (2025, May). SelfReplay: Adapting Self-Supervised Sensory Models via Adaptive Meta-Task Replay. *SenSys '25: Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems*. pp. 226 – 239. doi: 10.1145/3715014.3722066

Zhang, C., Cai, Y., Lin, G. and Shen, C. (2023, May). DeepEMD: Differentiable Earth Mover's Distance for Few-Shot Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45, pp.5632-5648. doi: 10.1109/TPAMI.2022.3217373.

Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A. & Kong, T. (2022). iBOT: Image BERT Pre-Training with Online Tokenizer. *The Tenth International Conference on Learning Representations (ICLR 2022)*. doi: 10.48550/arXiv.2111.07832.