LEVIS: A VISION TRANSFORMER FOR FAST COMBINATORIAL OPTIMIZATION OF IMAGING TECHNIQUES

Anonymous authors
Paper under double-blind review

ABSTRACT

The performance of a vision model depends greatly on the types of data on which it is trained. For applications from classifying objects to predicting the weather, models can be trained on a variety of image sensors and sampling patterns. Vision systems often face practical constraints (e.g. acquisition time, compute resources, energy, and memory consumption) that limit the number of sensors and samples. In such cases, obtaining and training the model on a subset of available modalities is beneficial. Although many iterative combinatorial optimization algorithms can find optimal sets of modalities, they are drastically slowed by the need to train the model to evaluate each proposed set. We introduce LeViS (Learned Vision System), a vision transformer that proficiently classifies images collected with any sampling pattern or sensor set *without retraining*. The predictive power of a set can of techniques can thus be quickly determined by evaluating the performance of LeViS on the set. We use LeViS with a variety of optimization algorithms (genetic algorithms, beam search, simulated annealing, sequential) to rapidly find optimal sets of satellite wavelength channels to classify local climate zones using So2Sat, and optimal pixel sampling patterns to classify handwritten digits and charaters using MNIST. Evaluating sets with LeViS instead of training new models enables optimization algorithms to find optimal sampling patterns and sensor sets up to 6800x faster.

1 INTRODUCTION

For many applications of computer vision, especially in scientific and industrial settings, many imaging techniques and sampling patterns are applicable. Microscopes, for example, can image hundreds of channels, using brightfield, phase contrast, fluorescent markers, and Raman spectroscopy to analyze the morphology and behavior of cells (Fischer et al., 2011; Pinkard et al., 2024; Bray et al., 2016; Zumbusch et al., 1999). Hyperspectral satellites image the Earth for years with hundreds of wavelength bands for agricultural, military, environmental, and civil applications (Zhu et al., 2020; Bhargava et al., 2024). Robotic perception systems commonly use RGB and IR cameras, ultrasound, lidar, radar, and haptics to map the environment (Liu et al., 2024).

Beyond choosing which imaging methods to use, vision systems must choose how to sample the world with each method. Microscopy methods such as confocal microscopy or two-photon microscopy, as well as macroscopic imaging methods such as lidar, steer the focus of a laser to build 3D images by scanning one point at a time. Robots must choose where to point their cameras. MRI machines measure points of the Fourier transform of the 3D image of a subject using radio waves (Zbontar et al., 2019).

Although we would like to use every imaging method to measure every point in space, we are often constrained in the amount of data we can image. For example, many microscopy methods require harsh chemicals or light, so live cells can only be imaged with a few imaging channels (Icha et al., 2017). When monitoring very fast microscopic dynamics such as neurons firing in brain tissue, scientists only have time to image only a few points in space since point-scanning the whole volume is too slow (Li et al., 2024). Hyperspectral satellites like NASA's Hyperspectral Infrared Imager satellite can measure 5.2TB/day, necessitating careful management of transmission bandwidth, power, and storage (Sun & Du, 2019). Mobile vision systems such as drones, robots, or self-driving cars have limited space, power, and computation, so the placement and number of cameras is essential.

Thus, to determine the constrained set of imaging techniques or sampling patterns that maximizes the performance of a task (e.g., classifciation, segmentation) when a model is trained on it, it is common to first collect and analyze an *anchor dataset* which contains subjects imaged with a wide variety of fully-sampled imaging methods. The optimal subset determined using the anchor dataset can then be deployed in the constrained environment. For example, fully-sampled MRI datasets can be used to determine fast scanning patterns (Zbontar et al., 2019). Biologists can image a single plate of cells with dozens of microscopy methods to determine the few methods to deploy in a drug discovery pipeline.

However, while exhaustive search is not practical due to rapid combinatorial explosion (˜126 trillion ways to choose 25 methods out of 50), there are a plethora of highly effective combinatorial optimization algorithms, such as genetic algorithms, simulated annealing, sequential selection, and beam search, which iteratively evaluate and propose sets to arrive at an optimum.

Unfortunately, evaluating each set of imaging methods is expensive, as vision models are usually trained to use a single set of channels; a CNN trained on RGB images will not work with x-ray images. For each proposed set of imaging techniques, the model needs to be retrained, making exploration of the design space computationally prohibitive. Many iterative combinatorial algorithms are thus impractical as they would require training hundreds of models.

In this work, we leverage the dimensional flexibility of the attention mechanism (Vaswani et al., 2023) to create a model that optimally classifies images regardless of the imaging techniques or spatial sampling pattern used to collect the image. To accomplish this, LeViS constructs patch tokens for each imaging method separately, encodes the tokens from each method into an compressed latent representation, then cross-attends to the latents to produce a prediciton. If LeViS is trained on a single patch pattern or set of channels, it is unable to generalize to different combinations of channels and sampling patterns at test time. To solve this, we randomly mask patches in the image or whole channels during training. This way LeViS to learns to opportunistically attend to all relevant information in the input and generate accurate predictions for arbitrary combinations of imaging techniques.

By encoding each imaging method separately and using cross-attention against the latents to generate a prediction, LeViS's computational requirements scale linearly with the size of the sampling pattern and the number of imaging methods used. LeViS can also cache the latent representations for each imaging method, allowing inference with different combinations of imaging methods on the anchor dataset by only running the decoder. Inference using cached latents in practice takes less than half the time of running the whole model, drastically accelerating combinatorial optimization.

2 Related Work

Vision transformers (ViT) have proved to be one of the most powerful architectures for various computer vision tasks (Dosovitskiy et al., 2020; Carion et al., 2020). Researchers have applied ViT for a variety of applications, including inferring gene function from multichannel microscopy, classifying deforestation using satellite imagery, and segmenting 3D MRI images, to name a few (Sivanandan et al., 2023; Kaselimi et al., 2023; Hatamizadeh et al., 2022).

A persistent issue with transformer architectures remains the quadratic scaling of the self-attention mechanism (Vaswani et al., 2023). This is especially problematic as the number of pixels scales quadratically with the height or width of 2D images and linearly with the number of imaging methods or channels, meaning that imaging applications with multi-modal imaging applications of transformers have massive computational requirements.

The Perceiver and Perceiver IO architectures use an asymmetric cross-attention mechanism to project large inputs into fixed-sized latents, then operate on the latents, to linear scaling with the size of the input (Jaegle et al., 2021; 2022). Perceivers support multimodal inputs, though modalities cannot be dynamically added or removed. Masked autoencoders completely decouple computational requirements from image size during training by selecting a random subset of patches from the image to operate on, but use the full image during inference (He et al., 2021).

Masked image encoding methods learn to process images corrupted by masking. Denoising autoencoders (DAE) have use a convolutional neural network (CNN) to denoise images masked by

noise (Vincent et al.). Context encoders and masked autoencoders leverage CNNs and transformers respectively to reconstruct arbitrary regions or patches removed from an image (Pathak et al., 2016; He et al., 2021). For hyperspectral imaging data, masked transformers have also been used to reconstruct random spatial-spectral blocks removed from satellite hyperspectral images (Scheibenreif et al., 2023). We note that spatial sampling of an image can easily be posed as masking all unsampled points in the image.

Models can also leverage masking in the channel dimension. Most machine learning models are purpose-built for a set of channels and do not function with only a subset of channels or with different channels. Strategies to make models channel-adaptive include training the model to expect all possible channels and inputting synthetic random data for unseen channels or tokenizing each channel of the input separately and adding a learned channel embedding to each token or learning convolutional weights and regularization parameters that can be applied to any channels (Shetab Boushehri et al., 2024; Bao et al., 2024; Chen et al., 2023).

Methods to choose sets of imaging techniques can be generally divided into three categories: analytical methods, searching methods, and embedded methods. Analytical methods analyze the statistics each imaging method directly to determine which to select. These include simple ranking methods that look at statistics of single pixel values such as variance, contrast, signal-to-noise ratio (SNR), and entropy to choose imaging methods or spectral bands (Liu et al., 2018; Chang et al., 1999), analyzing correlations between pixels of different imaging methods (MartÍnez-UsÓMartinez-Uso et al., 2007). Analytical methods do not take into consideration how combinations of channels perform with the model that analyzes them.

Searching methods iteratively propose a combination and test how well a model works with the combination. Many classical optimization algorithms are applicable to guide the search, including probabilistic search algorithms such as genetic algorithms and simulated annealing, heuristic search algorithms like beam search and hill climbing, and sequential search algorithms. Other more recent combinatorial optimization methods, such as using a neural solver Gao et al. (2025) or using combinatorial bayesian optimization Oh et al. (2019); Deshwal et al. (2023) are also applicable and potentially more sample efficient. The largest downside of searching methods is the expense of each evaluation, requiring retraining a model. Some methods of imaging technique optimization attempt to avoid retraining by observing the average effect of many random corruptions of each channel, but these techniques only give information about single-channel statistics, and require more than 100 evaluations of the whole dataset for each imaging technique to get reliable statistics.

Embedded methods have emerged as a promising technique to select imaging methods. In one strategy, candidate imaging channels are partitioned and ranked to isolate the most informative subset. These reduced bands are then passed through a hemispherical reflectance-based spatial filter and a 3D CNN, achieving highly effective classification (Phaneendra Kumar et al., 2024). Another method proposes a joint deep learning architecture composed of a constrained measurement learning network followed by a classifier, where the network directly learns to select bands that maximize task performance rather than relying on hand-crafted metrics (Ayna et al., 2023). Methods such as GFSNetwork employ temperature-controlled Gumbel-Sigmoid sampling to automatically determine informative subsets during training, offering a scalable solution that improves efficiency while preserving interpretability (Wydmański & Śmieja, 2025). Researchers have also analyzed the attention patterns of attention-based convolutional networks trained on all possible channels, then selected the subset of channels that were most attended to Ribalta Lorenzo et al. (2020).

3 METHOD

Given an input imaged with any collection of imaging methods or an arbitrary pixel sampling patterns, LeViS is designed to generate the best possible output for the data that it is given. This allows us to use LeViS as a fast evaluation function for the effectiveness of different sets of imaging methods or spatial sampling patterns without training a new model from scratch, greatly accelerating a wide variety of combinatorial optimization algorithms.

In practice, using any subset of channels or pixel-patterns in an image, LeViS can either (i) classify the input or (ii) reconstruct the original image. This dual function supports both **hypothesis-driven** optimization (e.g., optimizing classification performance for a specific task such as classification

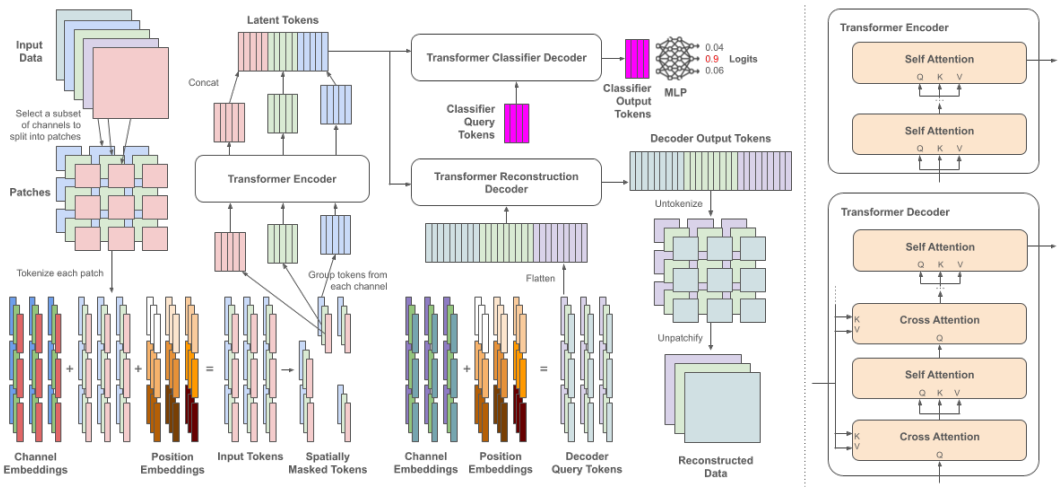

Figure 1: LeViS architecture overview

or segmentation) and **hypothesis-free** optimization (e.g., identifying sampling patterns maximally representative for the raw data).

3.1 LeViS Model Architecture

Channel Masking: Consider an input image x with dimensions $H \times W \times C$ from the anchor dataset, with spatial dimensions $H \times W$ and imaged using C imaging methods or channels. During training, a random number $C_{in} \in [1, C]$ is selected uniformly at random and C_{in} channels from the image are randomly selected yielding a $H \times W \times C_{in}$ image to be input into the model. This dropout procedure, introduced by Bao et al. (2024), regularizes the model to produce stable results no matter which set of imaging methods are inputted into the model.

Tokenizer: Each channel of the image is separately split into non-overlapping square patches of $P \times P$ pixels, yielding C_{in} sequences of $H//P \cdot W//P$ patches. A small convolutional neural network is applied to each patch to produce a token. As is standard in vision transformers, a sine-cosine position embedding is added to each token based on its position of origin(Dosovitskiy et al., 2020). A learned channel embedding is added to each token based on its channel of origin as in Bao et al. (2024) (Figure 1).

Spatial Masking: During training, a random number $n_{in} \in [1,\ H//P \cdot W//P]$ is selected uniformly at random, and all tokens in same n_{in} positions in each of the C_{in} sequences are deleted. This amounts to a randomly applied spatial mask on the image, similar to the masking procedure introduced in He et al. (2021).

Encoder: Tokens from each channel in the input are fed through the self-attention layers of the encoder separately to create a latent representation of each channel, which are concatenated and sent to the decoder. Thus, each token in the latent space arises from *a single channel*. Self-attention is applied to each channel separately, meaning that encoder computation scales linearly with the number of imaging methods used. Any combination of channels may be inputted, including single channels, so the latent representation of each channel must learn to contain all potentially relevant information for the decoder even if the information is redundant to other channels.

Classifier Decoder: A classification token and register tokens are passed through a series of alternating cross-attention layers (attending to the latent tokens) and self-attention layers. The classification token is fed into an MLP that produces class logits.

Reconstruction Decoder: Similarly to the classifier decoder, a set of learned query tokens, one for each patch in each channel, is fed through a series of cross-attention then self-attention layers. Each query token is then fed through a convolutional neural network decoder to produce a square patch,

and all patches are assembled into an output image. The reconstruction decoder can be used for self-supervised pre-training of LeViS or hypothesis-free discovery of sets.

3.2 Set Optimization

After training, LeViS can be used rapidly evaluate the predictive power of different combinations of imaging methods or pixel sampling patterns by applying the combination or pattern to the validation section of the anchor dataset and scoring the performance. Since LeViS is trained to extract whatever semantically relevant is in the input regardless of the channels or spatial patches it contains, the validation performance is an excellent scoring function for the performance of the combination or pattern. LeViS can be used as a rapid evaluator to accelerate almost any iterative combinatorial optimization algorithm. In this paper we demonstrate optimization of the set of k imaging methods or spatial patches of n available on the following algorithms.

Single-Shot In single-shot optimization, we take the individually best-performing k channels, and combine them into a set. This approach is highly efficient, but typically leads to lower accuracy. A key limitation is redundancy: features carrying similar information may be needlessly included. Conversely, features that perform poorly in isolation but complement each other when combined are likely to be overlooked.

Forward Selection This is a greedy, bottom-up feature selection procedure. It begins with an empty set of features and iteratively adds the feature that provides the best performance when combined with the previously selected features, continuing until k features have been included. An issue with this and similar approaches is that they do not take into account complementary sets. The best set of size $k-1$ is not necessarily a subset of k.

Backward Selection A greedy, top-down procedure. It begins with a full set of features, and iteratively removes the feature that would least decrease performance, until k features are left. This is similar to forward selection, with similar pitfalls - the ideal set for $k+1$ may not contain the ideal set for k.

Genetic Algorithms Begin with a random population of feature subsets, each of size k. At each generation, the performance of these subsets is evaluated, and the top-performing ones are retained. New subsets are produced by combining elements from two parent subsets (crossover) and occasionally altering features at random (mutation). Over successive generations, this process allows the population to evolve toward better solutions. The randomness of crossover and mutation helps the search escape local optima.

Simulated Annealing Start with an initial solution and a high "temperature," allowing the algorithm to accept both better and worse candidate solutions, encouraging exploration. As the temperature decreases, the probability of accepting worse solutions diminishes, making the search more greedy over time. This method moves past locally suboptimal feature sets to discover combinations that work well together.

Beam Search A more general forward selection, with parameter b, the beam width. At the first iteration, keep the best b single-feature sets. At each subsequent step, all possible single-feature extensions of these b sets are considered, and the best b among them are retained. After k steps, return the bestLarger beam widths allow the search to explore a broader set of candidates, increasing the chance of recovering near-optimal feature combinations that would be missed under a narrow beam, at the cost of higher computational complexity.

4 Experiments

We evaluate LeViS on two image classification benchmarks: So2Sat (Zhu et al., 2020) and MNIST (Lecun et al., 1998), demonstrating its effectiveness for rapid combinatorial optimization of imaging methods and sampling patterns, respectively. So2Sat was chosen for its diversity of imaging methods, with 18 different satellite synthetic aperture radar and multispectral optical imaging methods used to capture each image, and its small size, which is representative of a typical anchor dataset. Labels of local climate zones (e.g., compact high rise buildings, dense trees, water) were assigned by hand by a team of experts. MNIST was chosen for its clear spatial structure, making spatial sam-

pling a meaningful and explainable task, and small size, representative of a typical anchor dataset. Each image contains a handwritten arabic numeral, 0 through 9.

4.1 Optimizing Combinations of Imaging Methods

For So2Sat we trained LeViS with 4x4 pixel patches, 2 self-attention layers in the encoder, and 1 cross-attention layer and 4 self-attention layers in the classifer. We used the random split of So2Sat, training for 3000 epochs over 48 hours on a server with 2 Nvidia RTX 5090 GPUs, achieving 98.8% Top-1 accuracy. For comparison, we achieved 97.81% accuracy with ResNet50 which has 25.6 million parameters, Zhu et al. (2020) achieved 97.82% accuracy using ViT-S/8 with 21 million parameters, and Bao et al. (2024) report 99.10% accuracy using ChannelViT-S/8 with 21 million parameters.

Trivially, the validation accuracy that LeViS attains given a set of imaging methods provides a lower bound on the utility of that set, as LeViS can be used for that set out of the box. However, if we train a LeViS on a single set of imaging methods without masking, the validation accuracy has the potential to marginally improve over the base LeViS model since part of the base model's representational capacity is used to support channels not in the set.

To test whether the LeViS's average validation accuracy is representative of the true utility of the set, we first generate 100 random combinations of 5 and 10 methods. We train the base LeViS model on all possible channels with the standard random masking procedure. Then for each set of methods, we compare the validation performance of the base model with the base model fine-tuned for 10 epochs, the rough number of epochs of fine tuning needed for training accuracy to converge (Figure 2).

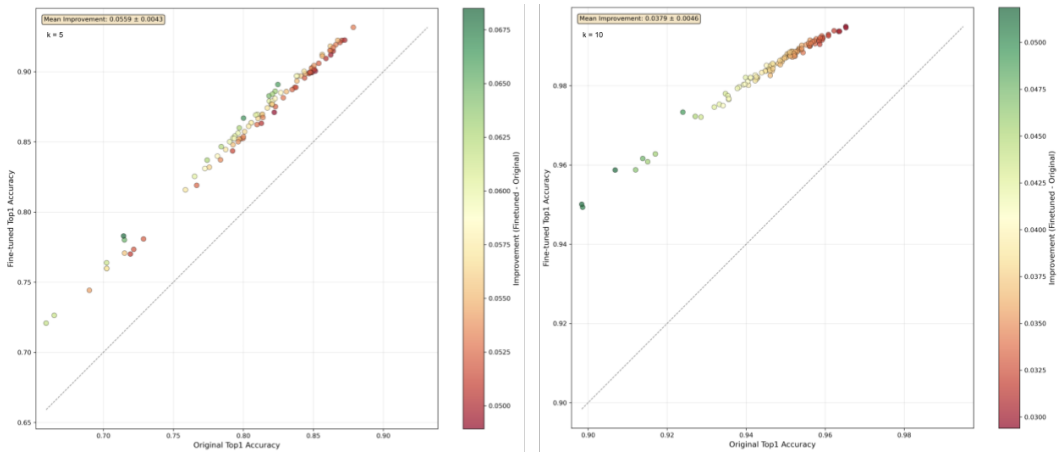

Figure 2: LeViS base accuracy vs fine-tuned accuracy on 100 random sets of 5 imaging methods (left) and 10 imaging methods (right)

Though we see modest increases in validation accuracies across all the combinations (avg +5.59% k=5, +3.79% k=10), the ordinality of scores produced by LeViS produces is largely equivalent to the scores of the fine-tuned model. This means that the validation scores generated by LeViS are almost always an accurate proxy to compare the predictive power of sets of imaging methods.

Leveraging LeViS as an ultrafast evaluator for different sets of imaging methods, we can quickly execute a variety of different iterative optimization algorithms. We first test a variety of optimization algorithms to find the optimal set of 5 channels from 18, with 8568 possible combinations we can quite easily compare the results to the optimum determined through exhaustive search (Table 1). We see that single-shot selection is extremely fast, as it simply picks the best-performing channels individually, but accuracy suffers significantly. Other methods find near-optimal solutions but take a few minutes. Notably, genetic algorithms and beam search perform the best, though less than a percent worse than the other methods. Most hyperparameters of the genetic algorithm are intuitive;

Table 1: Comparison of algorithms for selecting 5 channels of 18 from So2Sat

Algorithm	Channels	Best Top-1 Accuracy	Runtime (s)
Single Shot	[9, 11, 10, 17, 16]	0.7627	34.9
Simulated Annealing T=2 100 iters	[2, 4, 9, 13, 17]	0.8828	2087.4
Simulated Annealing T=1 100 iters	[2, 4, 9, 13, 17]	0.8828	2095.6
Simulated Annealing T=2 200 iters	[4, 5, 10, 12, 17]	0.8738	419.3
Simulated Annealing T=1 200 iters	[1, 5, 9, 13, 16]	0.8822	416.3
Genetic Algorithm pop=16 500 gens	[3, 4, 8, 12, 17]	0.8799	253.1
Genetic Algorithm pop=16 1000 gens	[3, 4, 8, 12, 17]	0.8799	160.3
Genetic Algorithm pop=4 500 gens	[3, 8, 10, 15, 16]	0.8722	91.8
Genetic Algorithm pop=4 20 gens	[3, 8, 10, 14, 16]	0.8716	51.0
Beam Search b=3	[1, 5, 9, 13, 17]	0.8860	410.7
Forward Selection	[9, 14, 5, 17, 1]	0.8827	159.3
Backward Selection	[3, 4, 8, 15, 17]	0.8789	370.2
Exhaustive search	[1, 5, 9, 13, 17]	0.8860	8938.5
Baseline	All channels	98.77	2.14

more generations and larger populations yield better results. Beam search, though the most effective, takes much longer than all other methods with little marginal benefit.

The optimization time trials were run on a desktop computer with a Nvidia RTX 4090 GPU and AMD CPU with 32 cores. To run inference on the entire So2Sat validation dataset without using cached latents, LeViS took 5.52 seconds and around 13 GB of VRAM with a batch size of 512. For the same task using cached latents, LeViS took 1.96 seconds and 1.09 GB of VRAM, saving considerable time and computational resources. For comparison, training the ResNet50 architecture He et al. (2015) which has a similar number of parameters to LeViS (25.6 million vs 17.9 million parameters, respectively) to 97.8% accuracy on the same computer took 2 hours and 28 minutes. This means that training a CNN model to evaluate a combination of imaging methods is more than 5000x slower than using LeViS.

4.2 Optimization of Sampling Patterns for Accuracy

For MNIST, we trained LeViS with 2x2 pixel patches, 2 self-attention layers in the encoder, and 2 cross-attention and 4-attention layers in the decoder, totaling 735 thousand parameters. We trained LeViS for 1000 epochs over 56 minutes on a desktop computer with a Nvidia RTX 4090 GPU, achieving a 98.78% accuracy with all patches.

Similar to our imaging method experiment, to test whether the LeViS's average validation accuracy is representative of the true utility of the set of patches inputted, we first generate 100 random combinations of 30 2x2 pixel patches out of the 28x28 image. We train the base LeViS model using the random masking procedure. Then for each set of patches, we compare the validation performance of the base model with the base model fine-tuned for 10 epochs (Figure 3).

Though we see a greater increase in validation accuracies across all the combinations (avg +9.25% k=30), the ordinality of scores produced by LeViS produces is remains largely equivalent to the scores of the fine-tuned model. This means that the validation scores generated by LeViS are almost always an accurate proxy to compare the predictive power of spatial sampling patterns.

We evaluated a variety of combinatorial optimization methods to optimize sampling patterns for MNIST digit classification using LeViS. We constrained all methods to use just 28 2x2 patches from the 196 patches available. This results in a much larger combinatorial space than with So2Sat, with upwards of trillions of possible combinations (Table 2). For optimizing sampling patterns with MNIST, there is a more clear tradeoff between runtime Top-1 Accuracy. Genetic algorithms punch above their weight, attaining a Top-1 accuracy within half of a percent of the best Top-1 accuracy with a small fraction of the runtime. Notably, LeViS only takes 490 milliseconds to process the entire MNIST validation dataset, saving 56 minutes of training time for each evaluation.

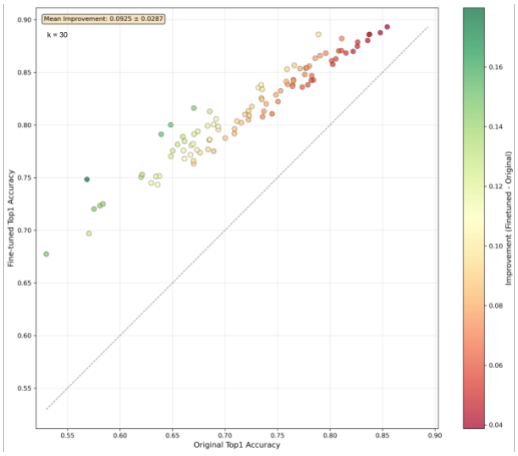

Figure 3: LeViS base accuracy vs fine-tuned accuracy on 100 random sets of 30 patches, MNIST

Table 2: Comparison of algorithms for selecting 30 2x2 pixel patches of 196 patches of MNIST

Algorithm	Best Top-1 Accuracy	Runtime (s)
Single Shot	0.8484	85.8
Simulated Annealing t=0.02 1000 iterations	0.9439	443.5
Simulated Annealing t=0.02 200 iterations	0.9176	89.9
Simulated Annealing t=0.01 1000 iterations	0.9494	446.6
Simulated Annealing t=0.01 200 iterations	0.9134	88.7
Simulated Annealing t=0.005 1000 iterations	0.9467	450.1
Simulated Annealing t=0.005 200 iterations	0.9391	88.3
Genetic Algorithm pop=16 500 gens mutation=0.05	0.9459	399.6
Genetic Algorithm pop=16 1000 gens mutation=0.05	0.9527	567.1
Genetic Algorithm pop=8 500 gens mutation=0.05	0.9432	273.4
Genetic Algorithm pop=8 1000 gens mutation=0.05	0.9475	327.9
Genetic Algorithm pop=4 500 gens mutation=0.05	0.9244	86.4
Genetic Algorithm pop=4 1000 gens mutation=0.05	0.9359	128.5
Beam Search width=3	0.9555	6968.4
Forward Selection	0.9541	2520.3
Backward Selection	0.9558	10866.5
Baseline	0.9878	0.48

4.3 Hypothesis-free Optimization of Sampling Patterns

Our goal with hypothesis-free optimization of sampling patterns is the search for the subset of pixel patches that contains the most information about the image as a whole. In practice, we search for the patch subset that can best reconstruct the whole image. We train LeViS on MNIST with 2x2 pixel patches, 2 self-attention layers in the encoder, and 2 cross-attention and 4-attention layers in the reconstruction decoder, totaling 1.1 million parameters. Training for 1000 epochs over 1.5 hours, we attain a final validation reconstruction MSE of 0.43. Using a genetic algorithm with population size 16 over 200 generations to optimize 30 patches, we obtain a reconstruction MSE of 0.715, which translates to a Top-1 accuracy of 79.2%.

5 Conclusion

In conclusion, our proposed method LeViS, vastly accelerates the combinatorial optimization of imaging methods and spatial sampling patterns for computer vision by eliminating the need to re-train a model to evaluate each proposed set. We demonstrate state of the art performance on two

image classification benchmarks: So2Sat (Zhu et al., 2020) and MNIST (Lecun et al., 1998), while reducing evaluation costs by several orders of magnitude. By serving as a fast, flexible evaluator, LeViS enables classical search strategies, such as genetic algorithms, beam search, and simulated annealing, to operate effectively in domains where exhaustive retraining is computationally infeasible. Beyond the tasks considered here, our framework opens the door to applying combinatorial optimization in broader scientific and industrial settings, including microscopy, robotics, and medical imaging. Future directions include extending LeViS to larger-scale multimodal datasets, integrating it with active data acquisition systems, and exploring its potential for real-time adaptive sensing. Overall, LeViS provides a practical foundation for rethinking how imaging pipelines are designed under real-world constraints.

REFERENCES

Cemre Omer Ayna, Robiulhossain Mdrafi, Qian Du, and Ali Cafer Gurbuz. Learning-Based Optimization of Hyperspectral Band Selection for Classification. *Remote Sensing*, 15(18):4460, January 2023. ISSN 2072-4292. doi: 10.3390/rs15184460. URL https://www.mdpi.com/2072-4292/15/18/4460. Publisher: Multidisciplinary Digital Publishing Institute.

Yujia Bao, Srinivasan Sivanandan, and Theofanis Karaletsos. Channel Vision Transformers: An Image Is Worth 1 x 16 x 16 Words, April 2024. URL http://arxiv.org/abs/2309.16108. arXiv:2309.16108 [cs].

Anuja Bhargava, Ashish Sachdeva, Kulbhushan Sharma, Mohammed H. Alsharif, Peerapong Uthansakul, and Monthippa Uthansakul. Hyperspectral imaging and its applications: A review. *Heliyon*, 10(12):e33208, June 2024. ISSN 2405-8440. doi: 10.1016/j.heliyon.2024.e33208. URL https://www.sciencedirect.com/science/article/pii/S2405844024092399.

Mark-Anthony Bray, Shantanu Singh, Han Han, Chadwick T. Davis, Blake Borgeson, Cathy Hartland, Maria Kost-Alimova, Sigrun M. Gustafsdottir, Christopher C. Gibson, and Anne E. Carpenter. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature Protocols*, 11(9):1757–1774, September 2016. ISSN 1750-2799. doi: 10.1038/nprot.2016.105. URL https://www.nature.com/articles/nprot.2016.105. Publisher: Nature Publishing Group.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers, May 2020. URL http://arxiv.org/abs/2005.12872. arXiv:2005.12872 [cs].

Chein-I. Chang, Qian Du, Tzu-Lung Sun, and M.L.G. Althouse. A joint band prioritization and band-decorrelation approach to band selection for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 37(6):2631–2641, November 1999. ISSN 1558-0644. doi: 10.1109/36.803411. URL https://ieeexplore.ieee.org/document/803411.

Zitong Chen, Chau Pham, Siqi Wang, Michael Doron, Nikita Moshkov, Bryan A. Plummer, and Juan C. Caicedo. CHAMMI: A benchmark for channel-adaptive models in microscopy imaging. November 2023. URL https://openreview.net/forum?id=Luc1bZLeMY.

Aryan Deshwal, Sebastian Ament, Maximilian Balandat, Eytan Bakshy, Janardhan Rao Doppa, and David Eriksson. Bayesian Optimization over High-Dimensional Combinatorial Spaces via Dictionary-based Embeddings. January 2023. URL https://openreview.net/forum?id=NIkjN2msy0.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. October 2020. URL https://openreview.net/forum?id=YicbFdNTTy.

Robert S. Fischer, Yicong Wu, Pakorn Kanchanawong, Hari Shroff, and Clare M. Waterman. Microscopy in 3D: a biologist's toolbox. *Trends in Cell Biology*, 21(12):682–691, December 2011. ISSN 0962-8924, 1879-3088. doi: 10.1016/j.tcb.2011.09.008. URL https://www.cell.

com/trends/cell-biology/abstract/S0962-8924(11)00198-X. Publisher: Elsevier.

Chengrui Gao, Haopu Shang, Ke Xue, and Chao Qian. Neural Solver Selection for Combinatorial Optimization, May 2025. URL http://arxiv.org/abs/2410.09693. arXiv:2410.09693 [math].

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, and Daguang Xu. UNETR: Transformers for 3D Medical Image Segmentation. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1748–1758, Waikoloa, HI, USA, January 2022. IEEE. ISBN 978-1-6654-0915-5. doi: 10.1109/WACV51458.2022.00181. URL https://ieeexplore.ieee.org/document/9706678/.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, December 2015. URL http://arxiv.org/abs/1512.03385. arXiv:1512.03385 [cs].

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners, December 2021. URL http://arxiv.org/abs/2111.06377. arXiv:2111.06377 [cs].

Jaroslav Icha, Michael Weber, Jennifer C. Waters, and Caren Norden. Phototoxicity in live fluorescence microscopy, and how to avoid it. *BioEssays*, 39(8): 1700003, 2017. ISSN 1521-1878. doi: 10.1002/bies.201700003. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bies.201700003. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bies.201700003.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General Perception with Iterative Attention, June 2021. URL http://arxiv.org/abs/2103.03206. arXiv:2103.03206 [cs].

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver IO: A General Architecture for Structured Inputs & Outputs, March 2022. URL http://arxiv.org/abs/2107.14795. arXiv:2107.14795 [cs].

Maria Kaselimi, Athanasios Voulodimos, Ioannis Daskalopoulos, Nikolaos Doulamis, and Anastasios Doulamis. A Vision Transformer Model for Convolution-Free Multilabel Classification of Satellite Imagery in Deforestation Monitoring. *IEEE Transactions on Neural Networks and Learning Systems*, 34(7):3299–3307, July 2023. ISSN 2162-2388. doi: 10.1109/TNNLS.2022.3144791. URL https://ieeexplore.ieee.org/document/9701667.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998. ISSN 1558-2256. doi: 10.1109/5.726791. URL https://ieeexplore.ieee.org/document/726791.

Haoyang Li, Quan Lu, Zhong Wang, Wenbo Zhang, Yu Wu, Yandong Sun, Yue Hu, Lehui Xiao, Dongping Zhong, Suhui Deng, and Shangguo Hou. Three-dimensional random-access confocal microscopy with 3D remote focusing system. *Communications Engineering*, 3(1):166, November 2024. ISSN 2731-3395. doi: 10.1038/s44172-024-00320-2. URL https://www.nature.com/articles/s44172-024-00320-2. Publisher: Nature Publishing Group.

Keng-Hao Liu, Shih-Yu Chen, Hung-Chang Chien, and Meng-Han Lu. Progressive Sample Processing of Band Selection for Hyperspectral Image Transmission. *Remote Sensing*, 10(3):367, March 2018. ISSN 2072-4292. doi: 10.3390/rs10030367. URL https://www.mdpi.com/2072-4292/10/3/367. Publisher: Multidisciplinary Digital Publishing Institute.

Yu Liu, Shuting Wang, Yuanlong Xie, Tifan Xiong, and Mingyuan Wu. A Review of Sensing Technologies for Indoor Autonomous Mobile Robots. *Sensors*, 24(4):1222, January 2024. ISSN 1424-8220. doi: 10.3390/s24041222. URL https://www.mdpi.com/1424-8220/24/4/1222. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

Adolfo MartÍnez-UsÓMartinez-Uso, Filiberto Pla, JosÉ MartÍnez Sotoca, and Pedro GarcÍa-Sevilla. Clustering-Based Hyperspectral Band Selection Using Information Measures. *IEEE Transactions on Geoscience and Remote Sensing*, 45(12):4158–4171, December 2007. ISSN 1558-0644. doi: 10.1109/TGRS.2007.904951. URL https://ieeexplore.ieee.org/document/4378560.

Changyong Oh, Jakub Tomczak, Efstratios Gavves, and Max Welling. Combinatorial Bayesian Optimization using the Graph Cartesian Product. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://papers.neurips.cc/paper_files/paper/2019/hash/2cb6b10338a7fc4117a80da24b582060-Abstract.html.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context Encoders: Feature Learning by Inpainting, November 2016. URL http://arxiv.org/abs/1604.07379. arXiv:1604.07379 [cs].

B. L. N. Phaneendra Kumar, Radhesyam Vaddi, Prabukumar Manoharan, L. Agilandeeswari, and V. Sangeetha. A new band selection framework for hyperspectral remote sensing image classification. *Scientific Reports*, 14(1):31836, December 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-83118-8. URL https://www.nature.com/articles/s41598-024-83118-8. Publisher: Nature Publishing Group.

Henry Pinkard, Cherry Liu, Fanice Nyatigo, Daniel A. Fletcher, and Laura Waller. The Berkeley Single Cell Computational Microscopy (BSCCM) Dataset, February 2024. URL http://arxiv.org/abs/2402.06191. arXiv:2402.06191 [cs].

Pablo Ribalta Lorenzo, Lukasz Tulczyjew, Michal Marcinkiewicz, and Jakub Nalepa. Hyperspectral Band Selection Using Attention-Based Convolutional Neural Networks. *IEEE Access*, 8:42384–42403, 2020. ISSN 2169-3536. doi: 10.1109/ACCESS.2020.2977454. URL https://ieeexplore.ieee.org/document/9019632.

Linus Scheibenreif, Michael Mommert, and Damian Borth. Masked Vision Transformers for Hyperspectral Image Classification. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2166–2176, Vancouver, BC, Canada, June 2023. IEEE. ISBN 979-8-3503-0249-3. doi: 10.1109/CVPRW59228.2023.00210. URL https://ieeexplore.ieee.org/document/10208365/.

Sayedali Shetab Boushehri, Aleksandra Kornivetc, Domink J.E. Winter, Salome Kazeminia, Katharina Essig, Fabian Schmich, and Carsten Marr. PXPermute reveals staining importance in multichannel imaging flow cytometry. *Cell Reports Methods*, 4(2):100715, February 2024. ISSN 2667-2375. doi: 10.1016/j.crmeth.2024.100715. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC10921034/.

Srinivasan Sivanandan, Bobby Leitmann, Eric Lubeck, Mohammad Muneeb Sultan, Panagiotis Stanitsas, Navpreet Ranu, Alexis Ewer, Jordan E. Mancuso, Zachary F. Phillips, Albert Kim, John W. Bisognano, John Cesarek, Fiorella Ruggiu, David Feldman, Daphne Koller, Eilon Sharon, Ajamete Kaykas, Max R. Salick, and Ci Chu. A Pooled Cell Painting CRISPR Screening Platform Enables de novo Inference of Gene Function by Self-supervised Deep Learning, August 2023. URL https://www.biorxiv.org/content/10.1101/2023.08.13.553051v3. Pages: 2023.08.13.553051 Section: New Results.

Weiwei Sun and Qian Du. Hyperspectral Band Selection: A Review. *IEEE Geoscience and Remote Sensing Magazine*, 7(2):118–139, June 2019. ISSN 2168-6831. doi: 10.1109/MGRS.2019.2911100. URL https://ieeexplore.ieee.org/document/8738051.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, August 2023. URL http://arxiv.org/abs/1706.03762. arXiv:1706.03762 [cs].

Pascal Vincent, Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Isabelle Lajoie, Yoshua Bengio, Yoshua Bengio, Pierre-Antoine Manzagol, and Pierre-Antoine Manzagol. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion.

Witold Wydmański and Marek Śmieja. GFSNetwork: Differentiable Feature Selection via Gumbel-Sigmoid Relaxation, March 2025. URL http://arxiv.org/abs/2503.13304. arXiv:2503.13304 [cs].

Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J. Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, Marc Parente, Krzysztof J. Geras, Joe Katsnelson, Hersh Chandarana, Zizhao Zhang, Michal Drozdzal, Adriana Romero, Michael Rabbat, Pascal Vincent, Nafissa Yakubova, James Pinkerton, Duo Wang, Erich Owens, C. Lawrence Zitnick, Michael P. Recht, Daniel K. Sodickson, and Yvonne W. Lui. fastMRI: An Open Dataset and Benchmarks for Accelerated MRI, December 2019. URL http://arxiv.org/abs/1811.08839. arXiv:1811.08839 [cs].

Xiao Xiang Zhu, Jingliang Hu, Chunping Qiu, Yilei Shi, Jian Kang, Lichao Mou, Hossein Bagheri, Matthias Haberle, Yuansheng Hua, Rong Huang, Lloyd Hughes, Hao Li, Yao Sun, Guichen Zhang, Shiyao Han, Michael Schmitt, and Yuanyuan Wang. So2Sat LCZ42: A Benchmark Data Set for the Classification of Global Local Climate Zones [Software and Data Sets]. *IEEE Geoscience and Remote Sensing Magazine*, 8(3):76–89, September 2020. ISSN 2168-6831. doi: 10.1109/MGRS.2020.2964708. URL https://ieeexplore.ieee.org/document/9014553.

Andreas Zumbusch, Gary R. Holtom, and X. Sunney Xie. Three-Dimensional Vibrational Imaging by Coherent Anti-Stokes Raman Scattering. *Physical Review Letters*, 82(20):4142–4145, May 1999. doi: 10.1103/PhysRevLett.82.4142. URL https://link.aps.org/doi/10.1103/PhysRevLett.82.4142. Publisher: American Physical Society.

A Appendix

A.1 Statement on use of AI tools

Anthropic Claude was used to aid and polish writing in this paper.