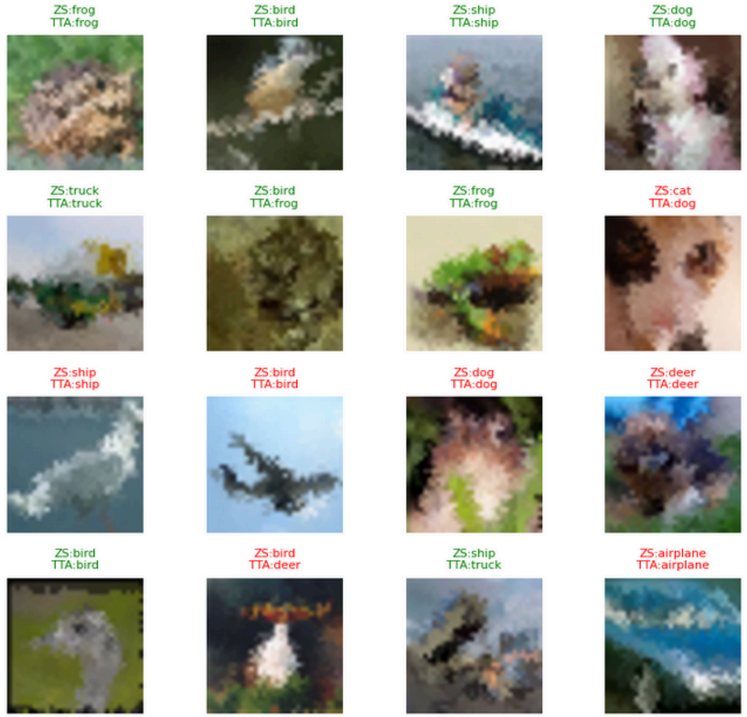
CLIPARTT: ADAPTATION OF CLIP TO NEW DOMAINS AT TEST TIME



MANAWADU D.N-220380J
MANAWADU M.D-220381M

Introduction

Self-Driving Car

AI-powered Robot

Pre-trained vision-language models like CLIP are not reliable under “domain shifts”

WHAT IS CLIP?

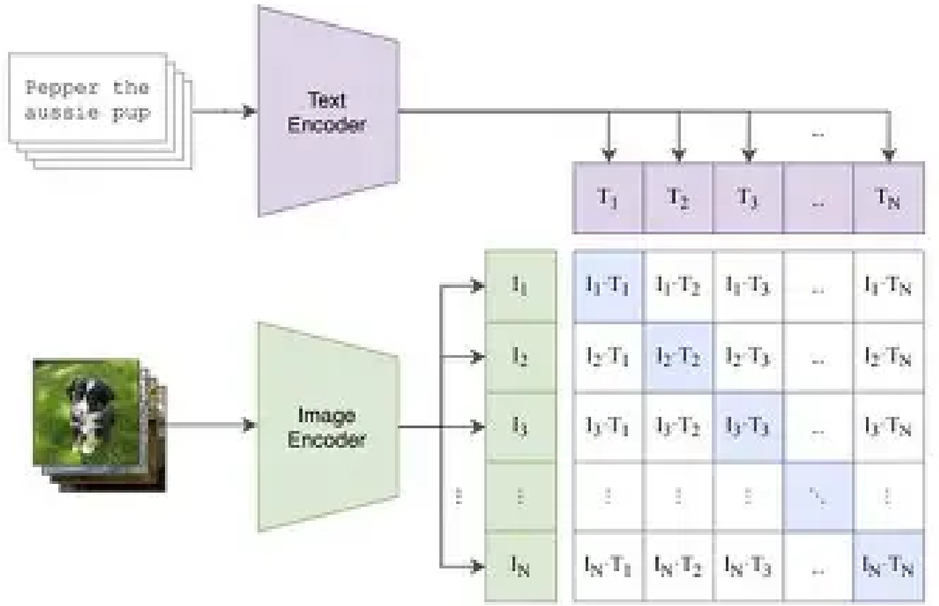
(Contrastive Language–Image Pretraining)

Text encoder (z_t)
‘a photo of a dog’

Visual encoder (z_v)
input image

Cosine similarity
$(z_v \cdot z_t / ||z_v|| ||z_t||)$

Highest similarity=predicted class



CLIP is trained to align matching image–text pairs and separate mismatched ones.

Background

TTA:-Test-Time Adaptation

- PTBN → Adjusts batch statistics at test time.
  - Limitation: Needs large batches.
- TENT → Refines affine parameters via entropy minimization.
  - Limitation: Relies on augmentations or large batches.
- CALIP → Uses features of the entire test set for hyperparameter search.
  - Limitation: Not scalable to very large datasets.
- TDA → Requires dataset-specific hyperparameter tuning.
  - Limitation: Poor generalization

Need: A method that is lightweight, simple, and effective.

CLIPARTT TEST-TIME ADAPTATION

Input batch of test images

CLIP Predictions (Top-K per image)

Create Multi-class Text Prompts e.g. "a photo of a truck or bus..."

Compute Similarities:
- Image-to-Image (Sv)
- Text-to-Text (St)

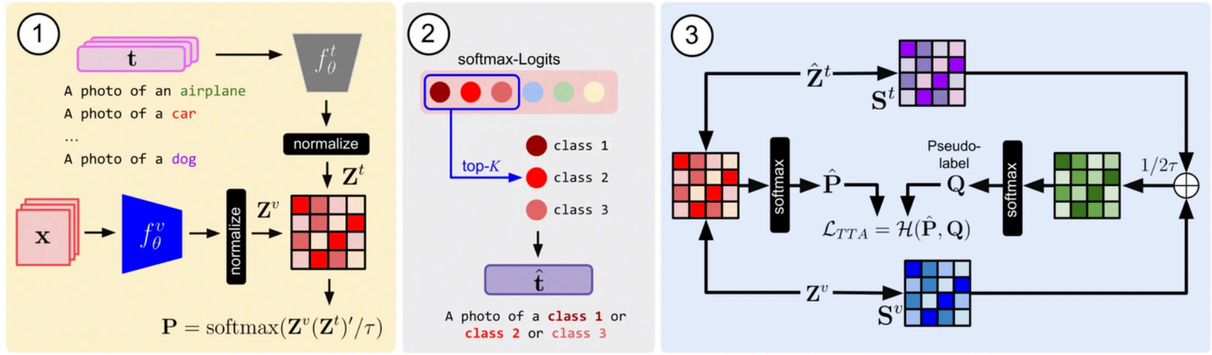
Combine → Pseudo-label Matrix Q (soft)

Predictions ĤP

Adaptation Loss:

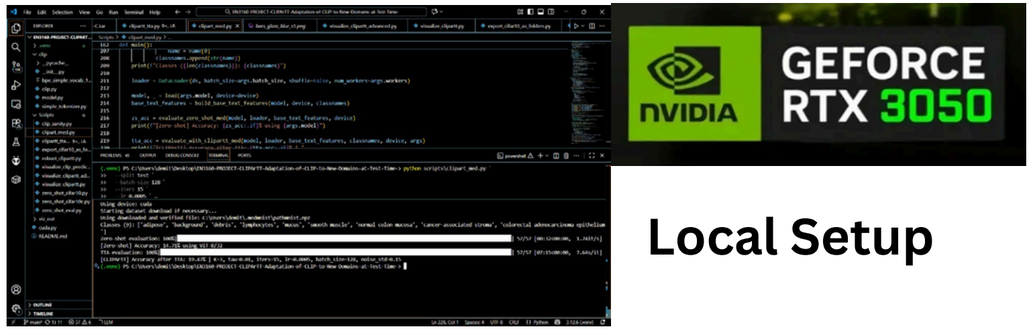
LTTA = CrossEntropy(Q , ĤP)

Update LayerNorm params of Visual Encoder.



IMPLEMENTATION SETUP

- Datasets: CIFAR-10, CIFAR-100, CIFAR-C, and ImageNet-C
- Model: CLIP (ViT-B/32 visual encoder)
- Batch size: 128
- Learning rate: 1×10−3
- Iterations per batch: 10
- Temperature (τ): 0.01
- Top-K classes: 3 (for instance-specific prompts)
- Optimizer: Adam
- Updated parameters: Only LayerNorm weights and biases in the visual encoder
- Hardware: CUDA-enabled GPU
- No retraining: Uses pretrained CLIP weights directly



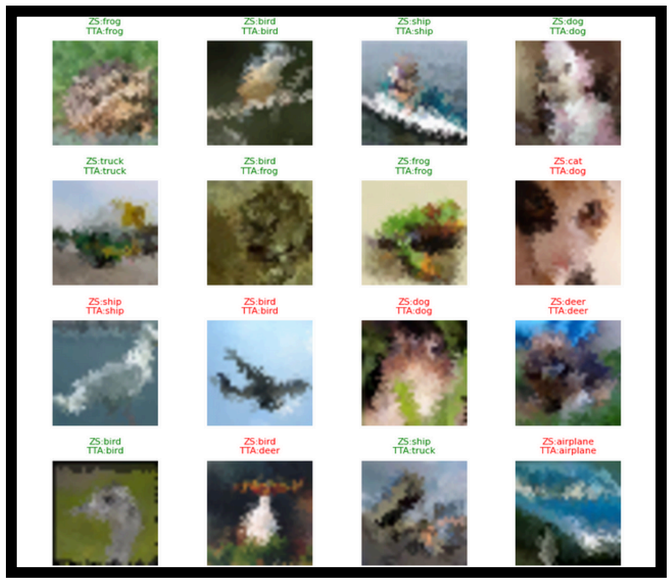
Local Setup

Cloud Setup

Replicated results

Performance comparison on CIFAR-10-C

Corruption Type	CLIP (Zero-shot)	CLIPArTT (After TTA)	Δ Improvement (↑)
Pixelate	48.45 %	66.58 %	+18.13 %
Impulse Noise	51.71 %	63.59 %	+11.88 %
Elastic Transform	53.19 %	64.99 %	+11.80 %
Contrast	61.98 %	73.38 %	+11.40 %




```
=== Running CLIPArTT on CIFAR100C (defocus_blur, severity=5) ===

Device: cuda
[Zero-shot] Accuracy: 42.19%
[CLIPArTT] Accuracy after TTA: 49.30% | dataset=cifar100c, corruption=defocus_blur, severity=5

=== Running CLIPArTT on CIFAR100C (glass_blur, severity=5) ===

Device: cuda
[Zero-shot] Accuracy: 19.44%
[CLIPArTT] Accuracy after TTA: 26.97% | dataset=cifar100c, corruption=glass_blur, severity=5

=== Running CLIPArTT on CIFAR100C (motion_blur, severity=5) ===

Device: cuda
```

Results

Performance omparison on CIFAR-100-C

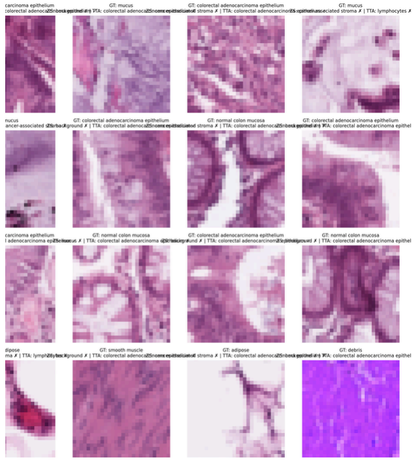
Corruption Type	CLIP (Zero-shot)	CLIPArTT (After TTA)	Δ Improvement (↑)
Defocus Blur	42.19 %	49.30 %	+7.11 %
Fog	41.14 %	49.10 %	+7.96 %
Brightness	57.08 %	63.53 %	+6.45 %
Gaussian Noise	18.61 %	25.23 %	+6.62 %

Corruption Type	CLIP (Zero-shot)	CLIPArTT (After TTA)	Δ Improvement (↑)
Impulse Noise	23.87 %	36.39 %	+12.52 %
Pixelate	23.11 %	35.91 %	+12.80 %
Contrast	34.38 %	45.19 %	+10.81 %
Defocus Blur	42.19 %	49.30 %	+7.11 %

NOVEL APPLICATION - MEDICAL IMAGES

Data Set-MedMNIST

A medical image dataset derived from real colorectal cancer slides.
Each 28×28 patch shows a distinct tissue type, such as adipose, stroma, or cancer epithelium.

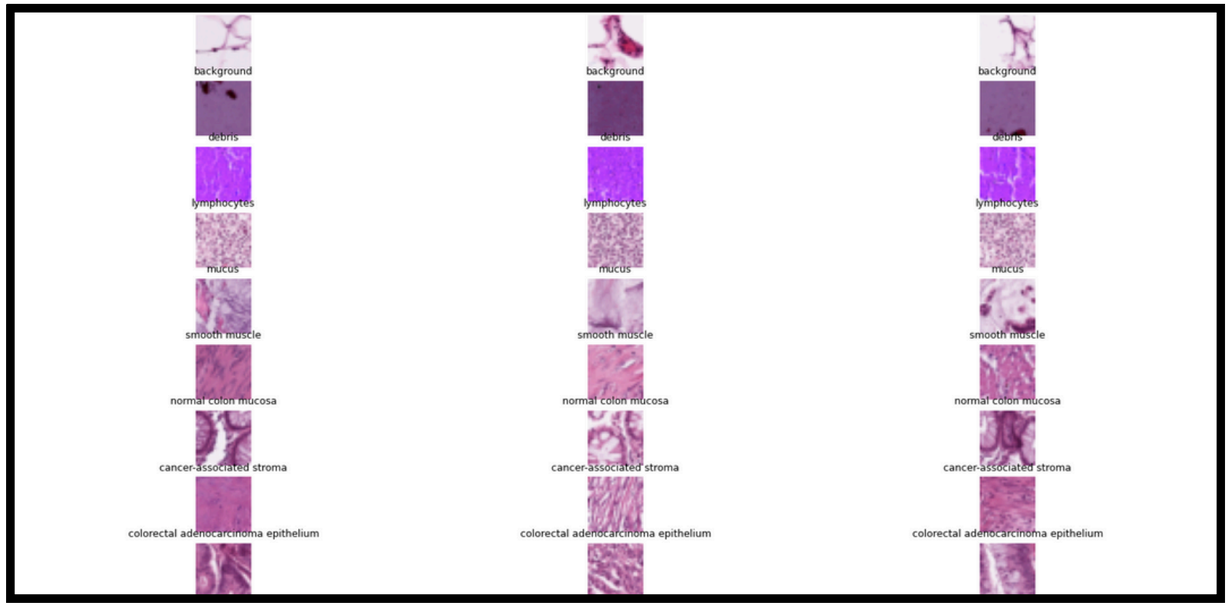


Clip Alone Performs
14.78%

With CLIPArTT
18.79%

MedMNIST2D

MedMNIST2D	Data Modality	Tasks (# Classes/Labels)	# Samples	# Training / Validation / Test
PathMNIST	Colon Pathology	Multi-Class (9)	107,180	89,996 / 10,004 / 7,180



Corrupted with Gaussian Noise

Why such a low accuracy on medical images ?

What is the significance of using this approach ?

```
Zero-shot evaluation: 100%|
                          | 57/57 [00:39<00:00,  1.44it/s]
[Zero-shot] Accuracy: 14.78% using ViT-B/32
TTA evaluation: 100%|
                          | 57/57 [07:26<00:00,  7.83s/it]
[CLIPArTT] Accuracy after TTA: 18.79% | K=3, tau=0.01, iters=15, lr=0.0005, batch_size=12
8, noise_std=0.15
(.venv) PS C:\Users\demit\Desktop\EN3160-PROJECT-CLIPArTT-Adaptation-of-CLIP-to-New-Domai
ns-at-Test-Time->
```

Conclusion

- CLIP is strong in zero-shot learning but weak under domain shifts.
- CLIPArTT adapts at test time → no retraining needed.
- Uses multi-class prompts + sample similarities for robustness.
- Only updates normalization layers → efficient.
- Can be introduced to new domains.