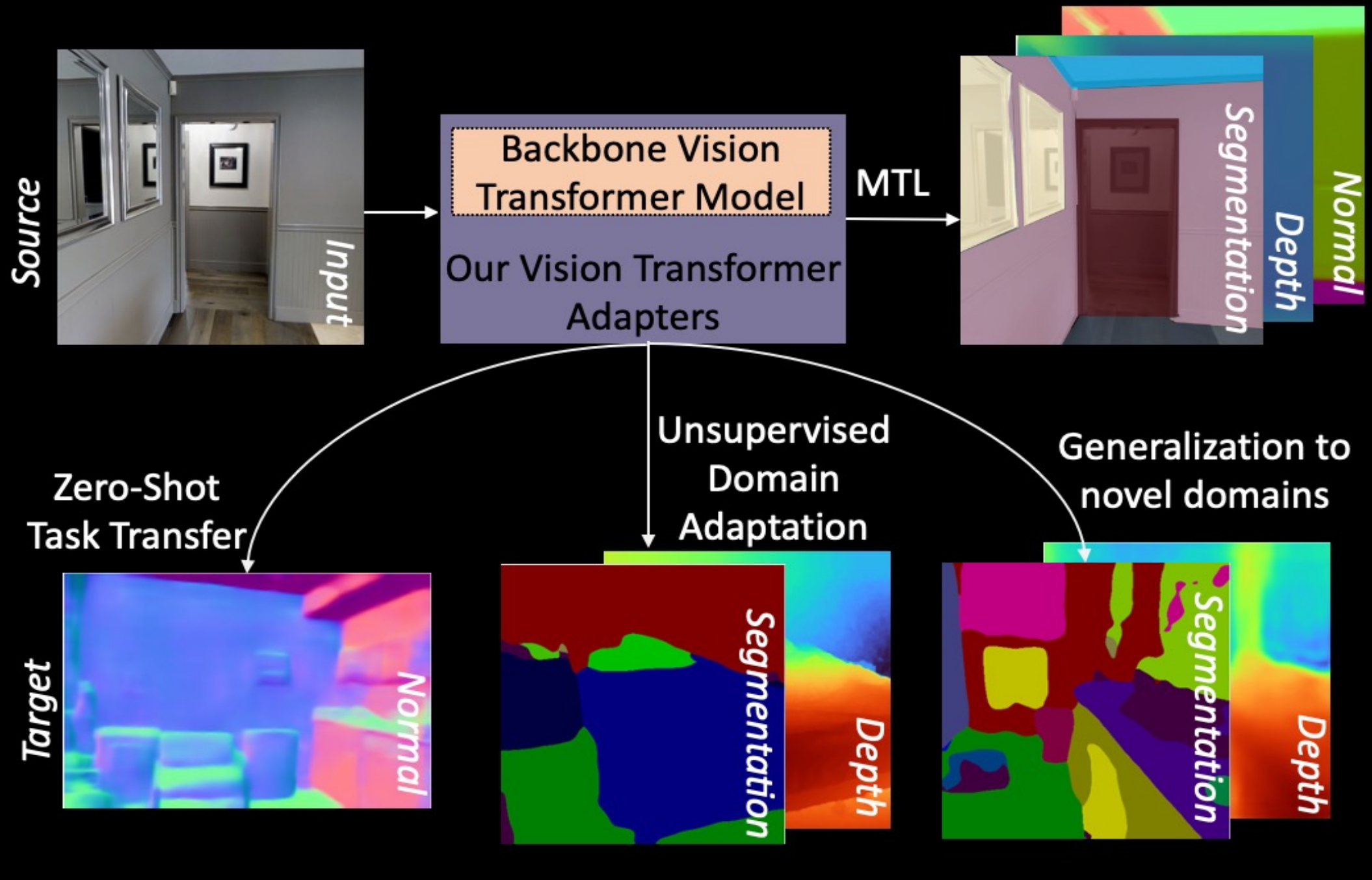
EPFL

ICCV23
PARIS

Vision Transformer Adapters for Generalizable Multitask Learning

Deblina Bhattacharjee, Sabine Süsstrunk, and Mathieu Salzmann

IVRL and CVLab, EPFL, Switzerland


Source
Input
Backbone Vision Transformer Model
Our Vision Transformer Adapters
MTL
Normal
Depth
Segmentation
Zero-Shot Task Transfer
Target
Normal
Unsupervised Domain Adaptation
Segmentation
Depth
Generalization to novel domains
Segmentation
Depth

Method Overview

H xWx 3

$Batch - size = 128$

TROA input: Encoded batch embeddings

TROA

T1 T2 T1 T2 T1 T15 T18 T1 T2

Stage 1 $\frac{H}{4}$ x $\frac{W}{4}$ x 192

Stage 2 $\frac{H}{8}$ x $\frac{W}{8}$ x 384

Stage 3 $\frac{H}{16}$ x $\frac{W}{16}$ x 768

Stage 4 $\frac{H}{32}$ x $\frac{W}{32}$ x 1536

Task head 1

Task head n

Task head N

$\frac{H}{32}$ x $\frac{W}{32}$ x 1536 $\frac{H}{16}$ x $\frac{W}{16}$ x 768 $\frac{H}{8}$ x $\frac{W}{8}$ x 384 $\frac{H}{4}$ x $\frac{W}{4}$ x 192

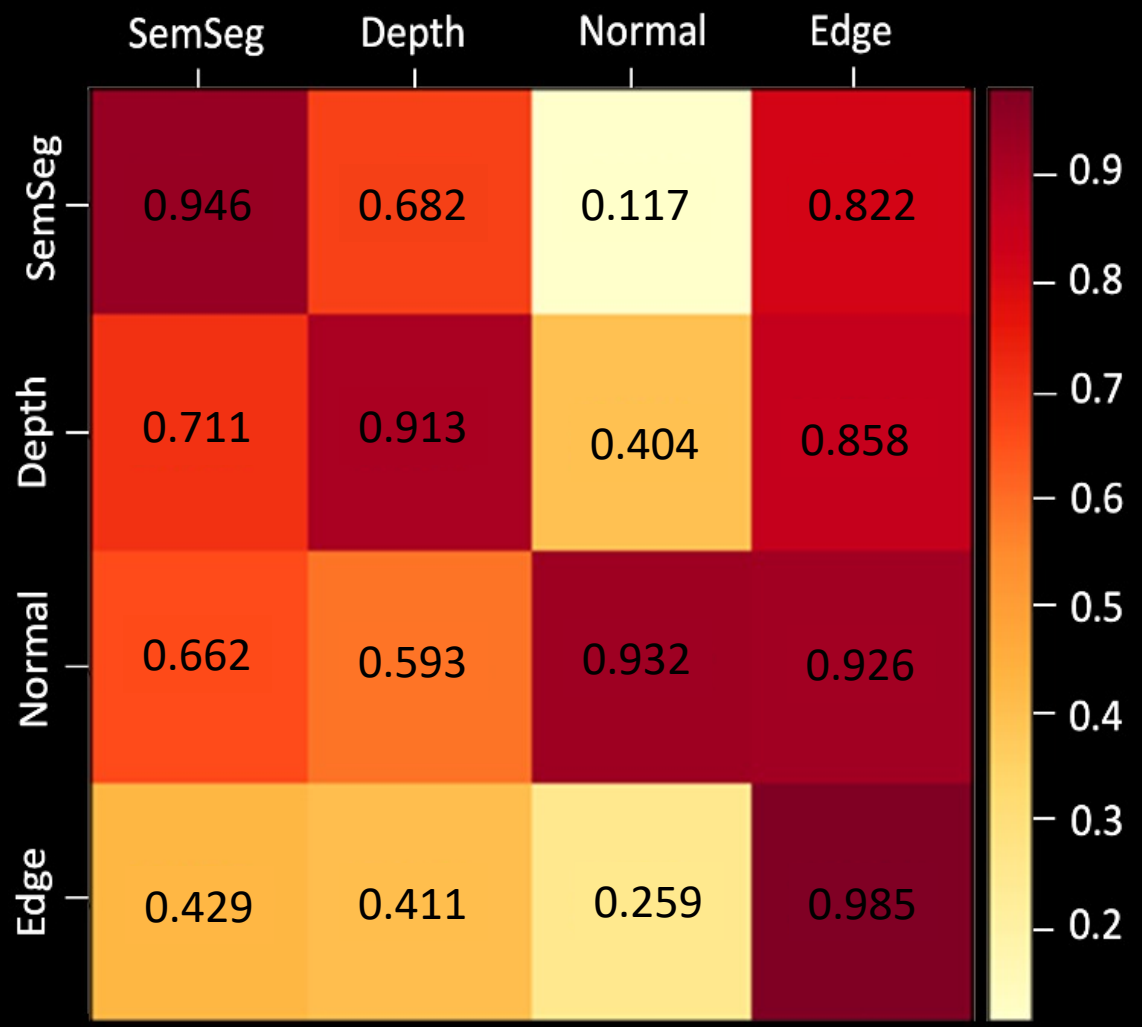
Legend

- Frozen Swin encoder
- Swin decoders trained
- Vision Adapter trained
- Task heads trained
- TROA trained
- Skip connection

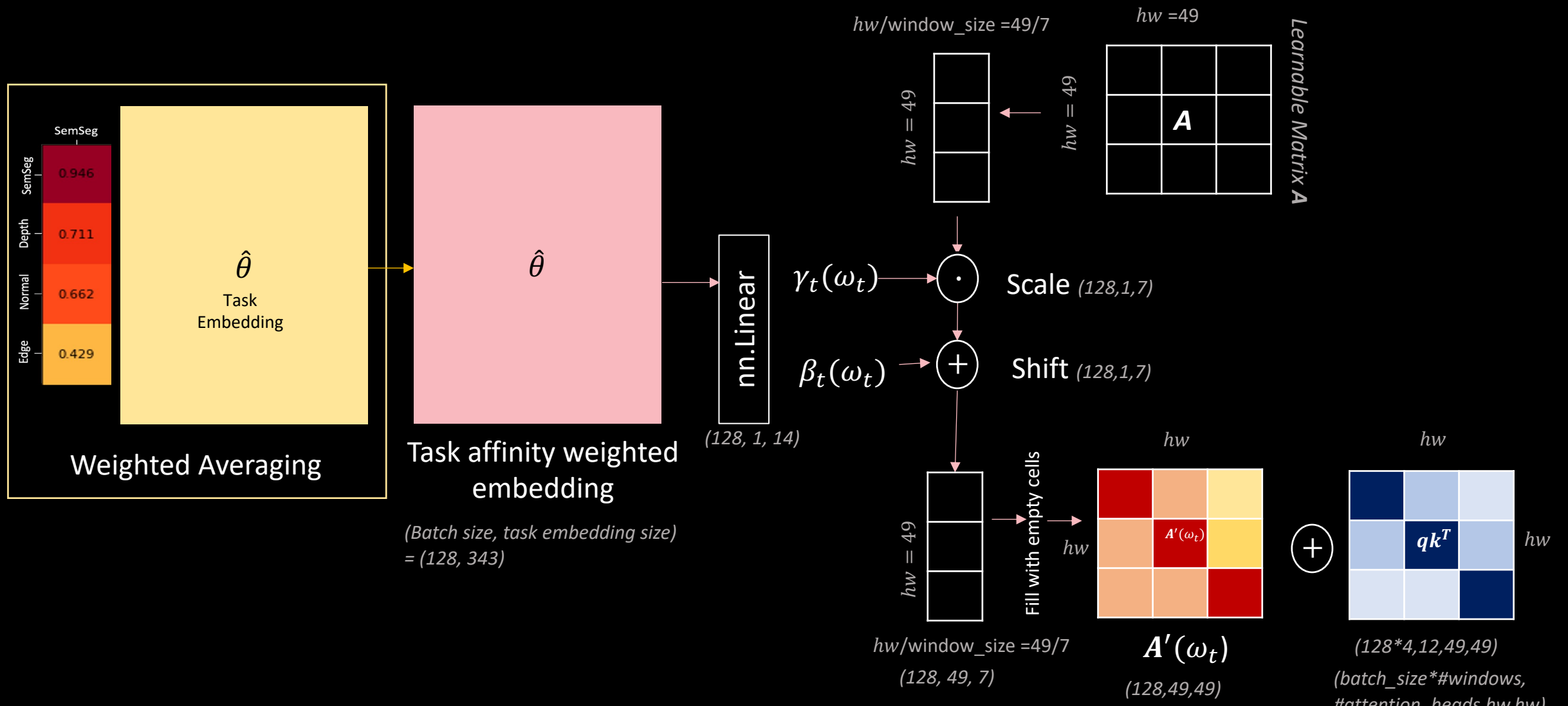
Motivation: Learning generalizable task affinities in a parameter-efficient way

Task Representation Optimization Algorithm (TROA)

Motivation: Finds gradient-based task affinities



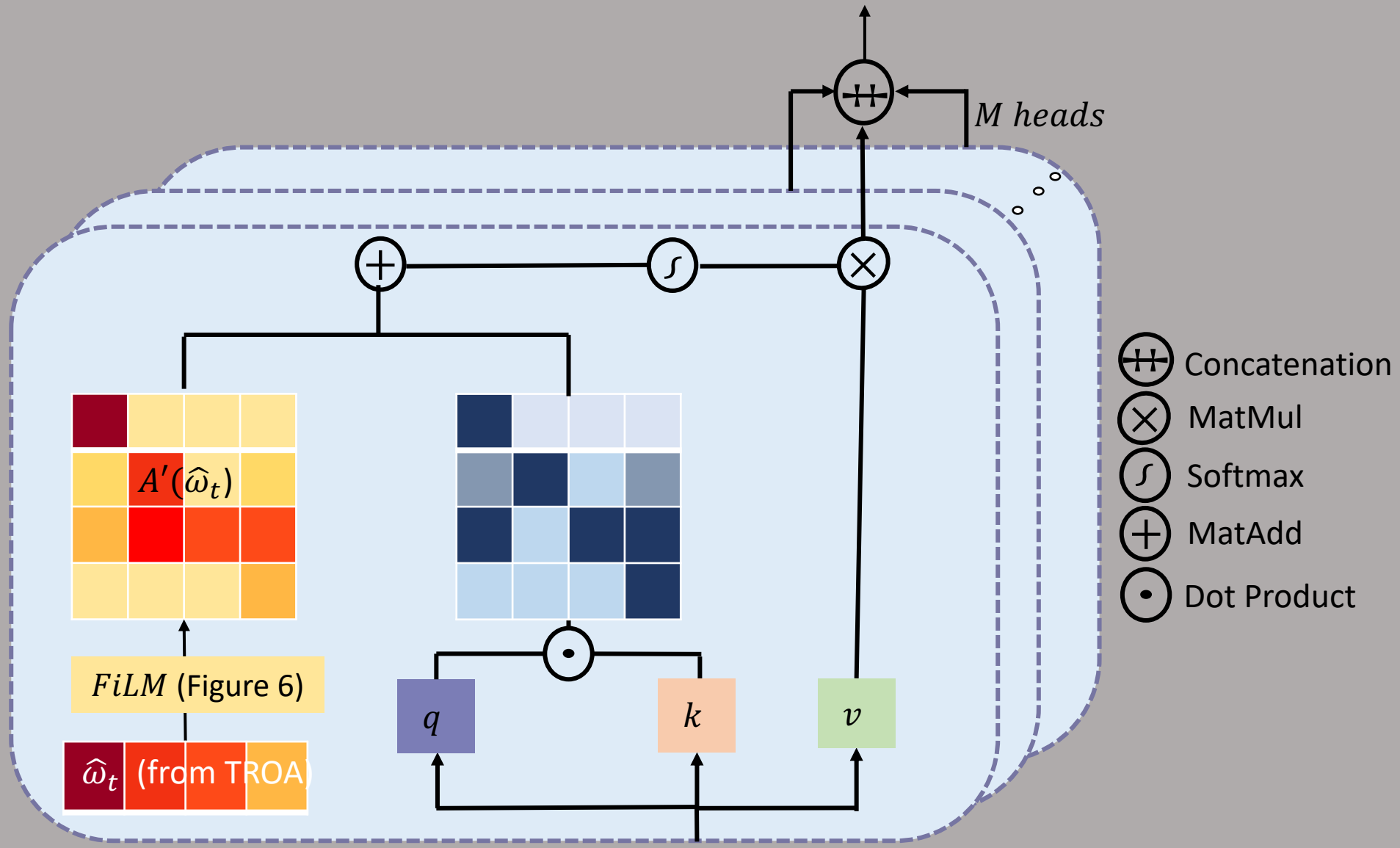
Feature-wise Linear Modulation



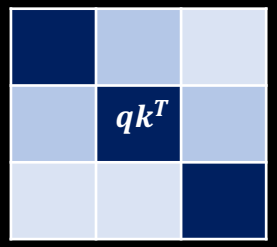
Perez et.al, MILA, published at AAAI 2018

Task-Adapted Attention (TAA)

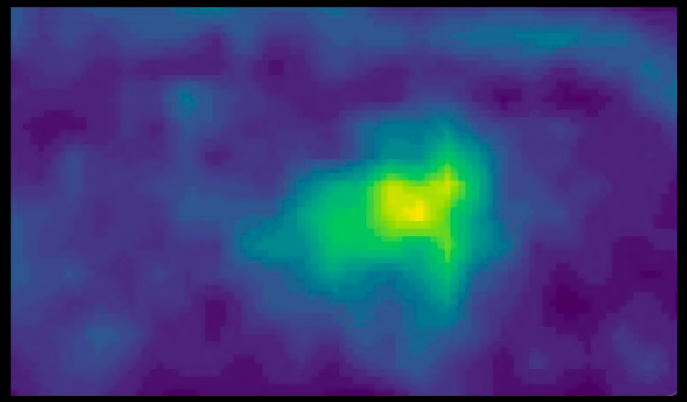
Motivation: Combines task affinities with image attention



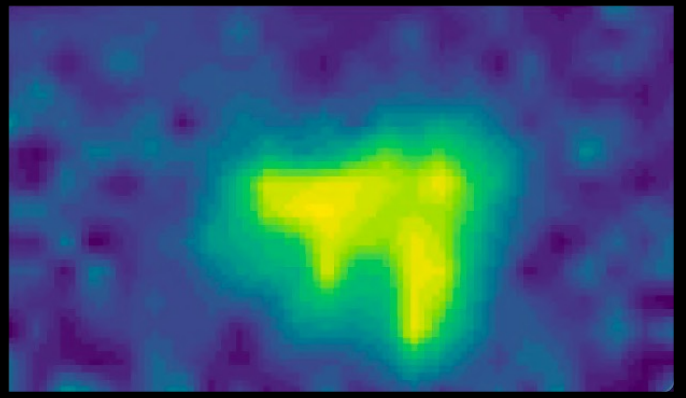
Task-Adapted Attention (TAA)



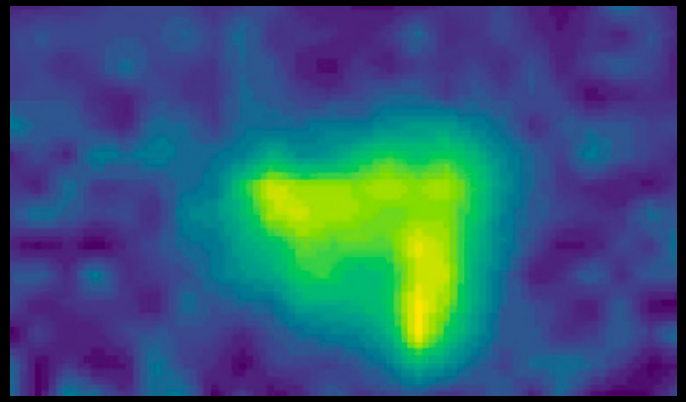
Example Image



Regular Self-Attention



SemSeg-TAA



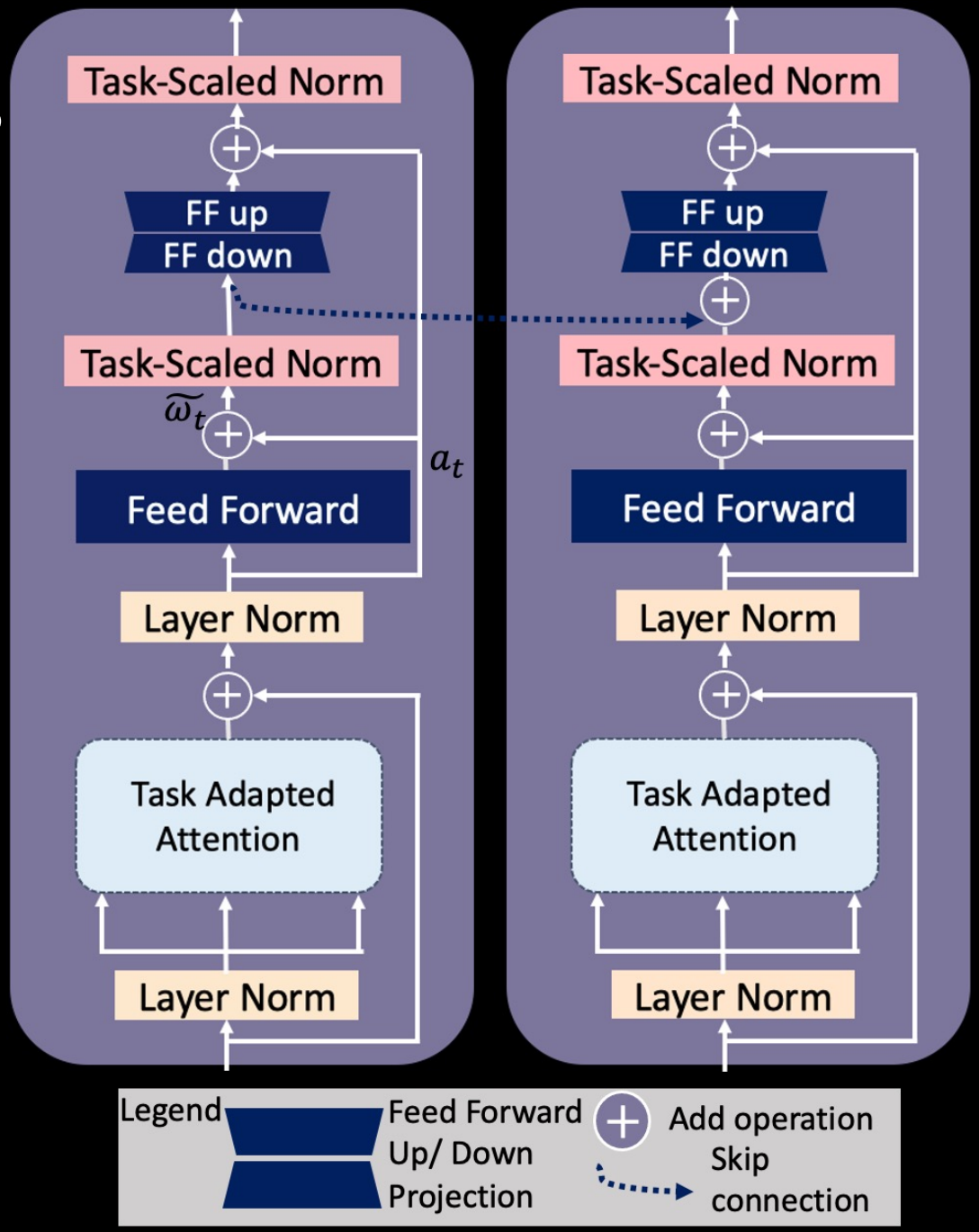
Depth-TAA

Vision Transformer Adapters

Motivation: Learns to perform multitasking in a parameter-efficient way; learns transferrable task affinities

To match the scales of different tasks

Normalizes across all features, more stable



Task-Scaled Norm

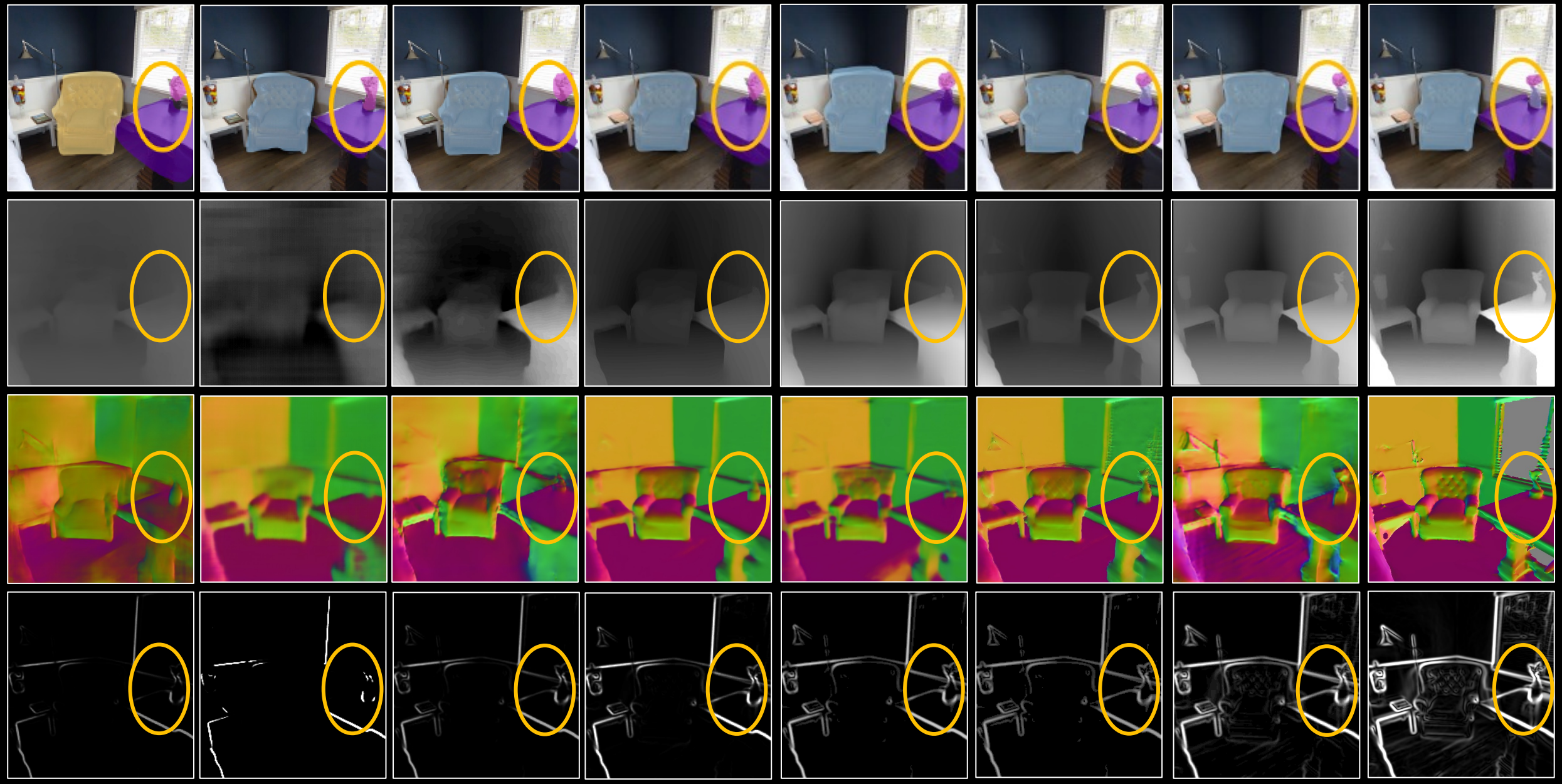
Motivation: TSN balances the different scales of the tasks. Balancing the task scales is necessary to avoid learning interference in MTL.

$$TSN_t = \frac{1}{\sigma}\left(a_t - \mu\right)\hat{\gamma}_t\left(\tilde{\omega}_t\right) + \beta_t\left(\tilde{\omega}_t\right)$$

$$\hat{\gamma}_t\left(\tilde{\omega}_t\right) = \gamma'\,\gamma_t\left(\tilde{\omega}_t\right) + \beta'$$

Swin's layer norm weights (for initialization)

Swin's Layer Norm: Fixed
Ours: Trained, based on the equation above



STL | MTL-baseline (CNN) | Consistency (CNN) | XTAM (CNN) | TAWT + MTL-baseline (CNN) | MulT (Transformer) | Adapters (transformer) | Ground-truth

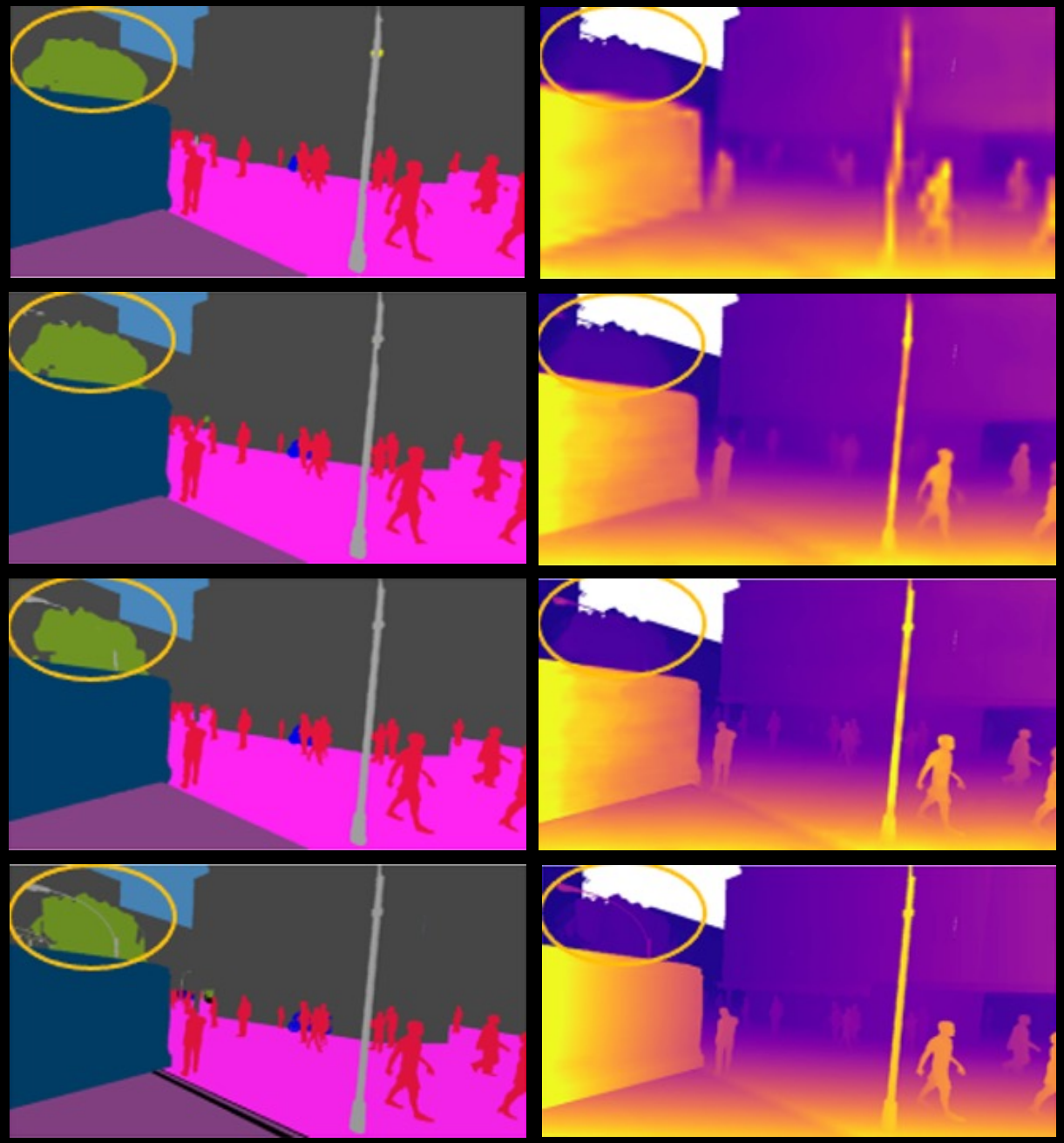
Zero-shot Task Transfer

TAWT + MTL-baseline

MulT

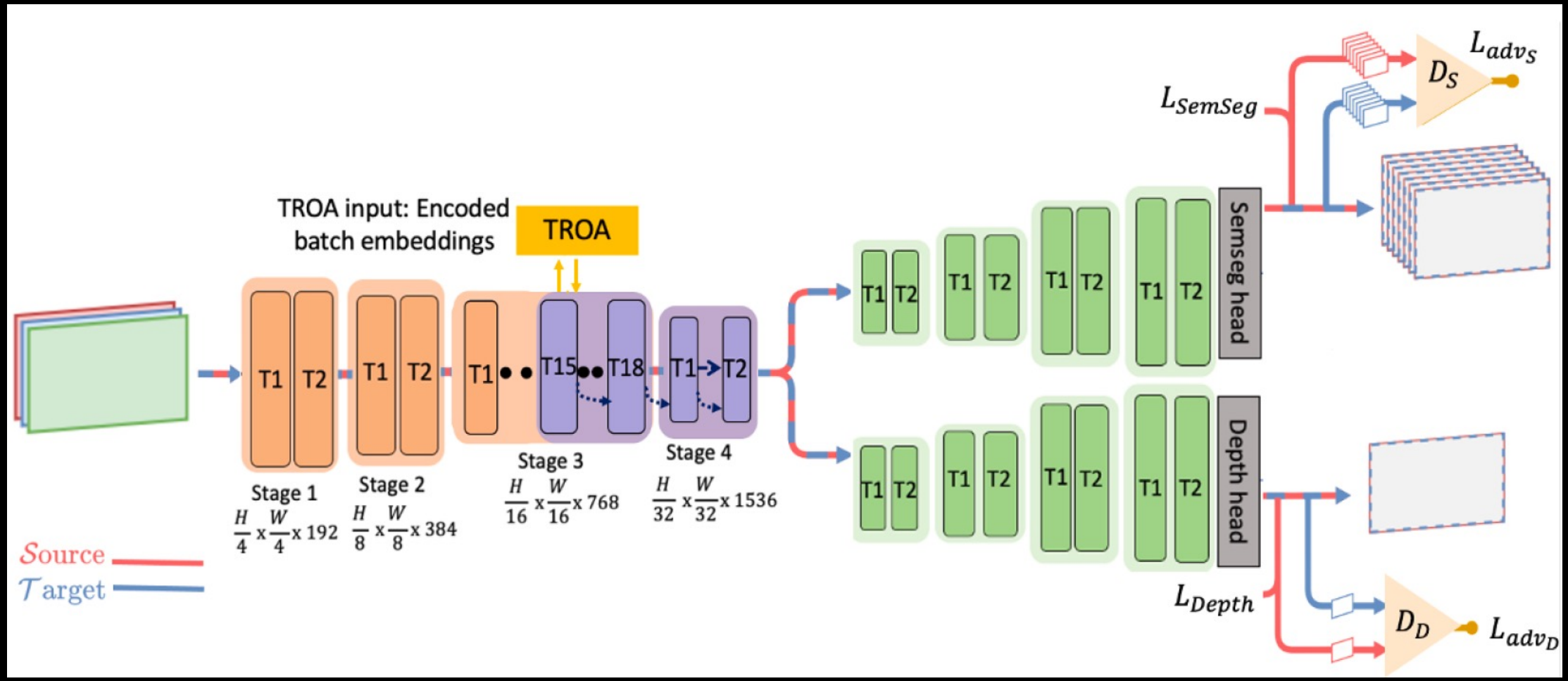
Adapters

Ground-truth

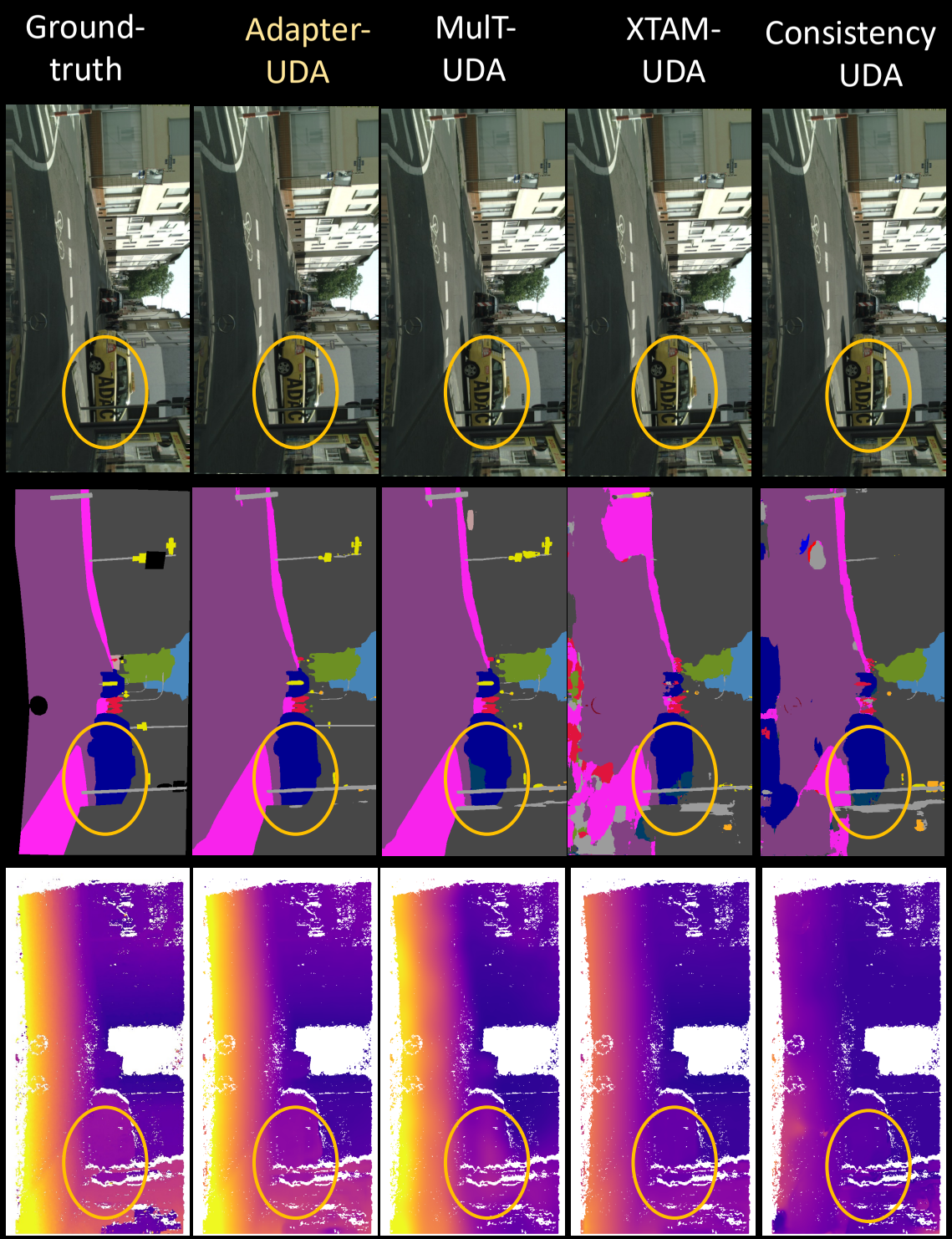


Normal is the zero-shot task

Unsupervised Domain Adaptation Method



Unsupervised Domain Adaptation



TEAM

Swiss National Science Foundation

FNSNF
FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION

EPFL

Thank you!

ICCV23
PARIS

This work was supported by the Swiss National Science Foundation via the **Sinergia grant CRSII5−180359**.