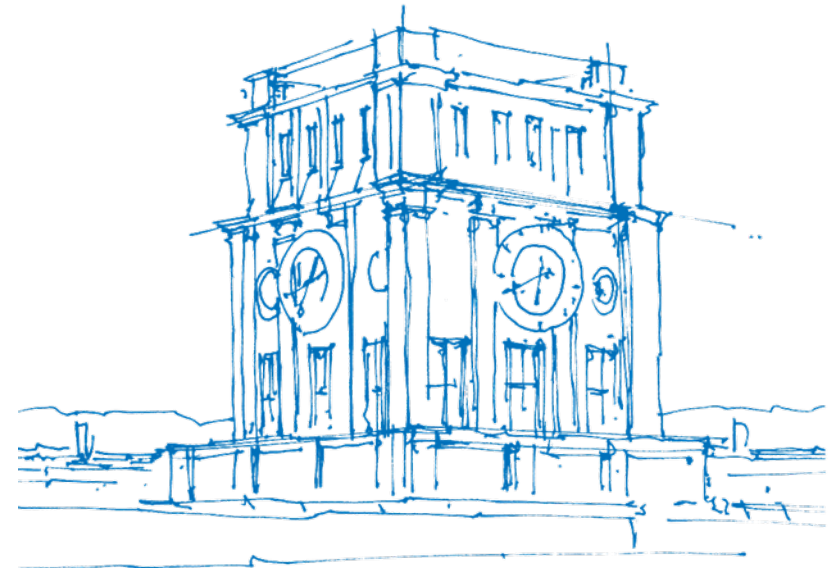


# Seminar: Foundation Models for Computer Vision

Tarun Yenamandra, Dominik Schnaus  
Computer Vision Group  
Technical University of Munich



*TUM Uhrenturm*

# How can I access these slides?

- **Option 1 (preferred):** seminar web page
  - <https://cvg.cit.tum.de/teaching/ws2024/fmcv>
  - Password for the materials page: fmcv24neu
  - Material page will go online after this pre-meeting
- **Option 2:** contact organizers
  - [fmcv-ws24@vision.in.tum.de](mailto:fmcv-ws24@vision.in.tum.de)
  - **Only use this option if you forgot the password**

# Outline

- General Information
  - About the Seminar
  - Registration
- Possible Topics
  - Self-Supervised Representation Learning in Computer Vision
  - Multi-Modal 2D Foundation Models
  - 3D Foundation Models
- Questions

# Outline

- General Information
  - About the Seminar
  - Registration
- Possible Topics
  - Self-Supervised Representation Learning in Computer Vision
  - Multi-Modal 2D Foundation Models
  - 3D Foundation Models
- Questions

# How is the seminar organized?

- Seminar meetings: Talks and discussion
  - **Block course**, course schedule will be announced later
  - Location: In presence (if required also virtual, on Zoom)
  - In case of special circumstances please let us know and we will find a solution
  - Each session will consist of two talks which are held in English
  - **Attendance is mandatory!**
- Talk preparation / contact with supervisor
  - Schedule meetings with your supervisor
  - Four weeks before talk: meet supervisor for questions (optional, but recommended)
  - Two weeks before talk: meet supervisor to go through slides (optional, but recommended)
  - One week before talk: send slides to your supervisor (mandatory)
  - Two weeks after talk: submit your report via email (mandatory)

# What about the presentation?

- General set-up:
  - Duration: 20–25 minutes talk + 10–15 minutes discussion
  - Make sure to finish on time - not too early and not too late!
  - Rule of thumb: 1–2 minutes per slide → 10–20 slides
  - Do not put too much information on the slides!
- Recommended structure (talk):
  - Introduction
  - Overview / Outline
  - Method description
  - Experiments and results
  - Personal comments
  - Summary

# What about the discussion after each talk?

- Discussion afterwards **will** influence your grade
- Ask questions!
- There are **no** stupid questions!

# What about the final report?

- General set-up:
  - Use  $\text{\LaTeX}$  template provided on web page
  - Length: 3-4 pages
  - Send final report as pdf by email to `fmcv-ws24@vision.in.tum.de`
  - Submission deadline: **Two weeks after talk**
- Recommended structure (main text only):
  - Introduction
  - Method description
  - Experiments and results
  - Discussion of results
  - Summary



# Outline

- General Information
  - About the Seminar
  - Registration
- Possible Topics
  - Self-Supervised Representation Learning in Computer Vision
  - Multi-Modal 2D Foundation Models
  - 3D Foundation Models
- Questions

# How do you register for the seminar?

- **Step 1:** Official registration via TUM matching system
  - Go to <https://matching.in.tum.de>
  - Register for seminar with the title *Foundational Models for 2D and 3D Computer Vision*
- **Step 2:** Personal registration via email
  - In the list of papers on the web page, select your three favorites
  - Write an email ranking these three favorites to the seminar email address
  - Email subject: “[FMCV] application [your name]”
  - Include information about related lectures / courses you have taken so far
  - We do **not** need your CV or a motivation letter!
  - Registrations without email / emails with missing information will be ignored!
- **Deadline** for both registrations: July 25<sup>th</sup>, 2024

# How do you register for the seminar?

Example registration email:

To:	fmcv-ws23@cvpr.in.tum.de
Cc:	
Subject:	[FMCV] application [your name]

Hi Tarun and Dominik,

I would like to present one of the following papers:

1. Paper A
2. Paper B
3. Paper C

In the past, I have taken these related courses:

- CV2 (winter 22)
- I2DL (summer 23)

Best,

# How do we select candidates and assign papers?

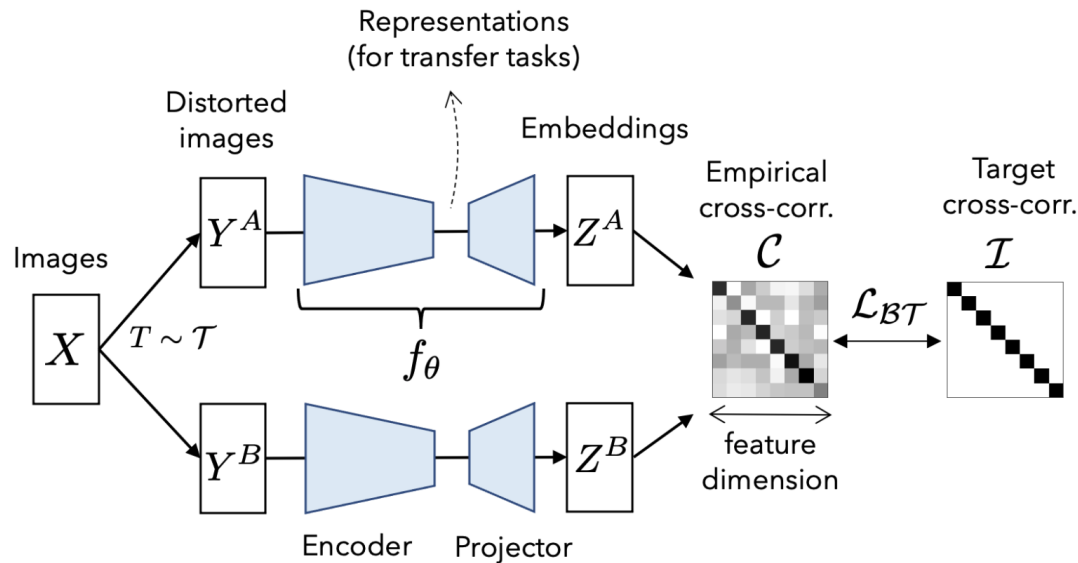
- Candidate selection
  - Only students registered in the matching system **AND** emails containing all required information will be considered
  - Among students meeting the formal criteria, selection will be random
  - Note that if you have not taken any related course, you must be willing to invest a lot of work to learn the required basics
  - You will get notified by the matching system about the decision
- Paper assignment
  - A formal paper list will be published on the materials page in the next weeks
  - Papers are assigned after the participant list is finalized
  - We give our best to accommodate your preference list in the assignment

# Outline

- General Information
  - About the Seminar
  - Registration
- Possible Topics
  - Self-Supervised Representation Learning in Computer Vision
  - Multi-Modal 2D Foundation Models
  - 3D Foundation Models
- Questions

# Contrastive Methods

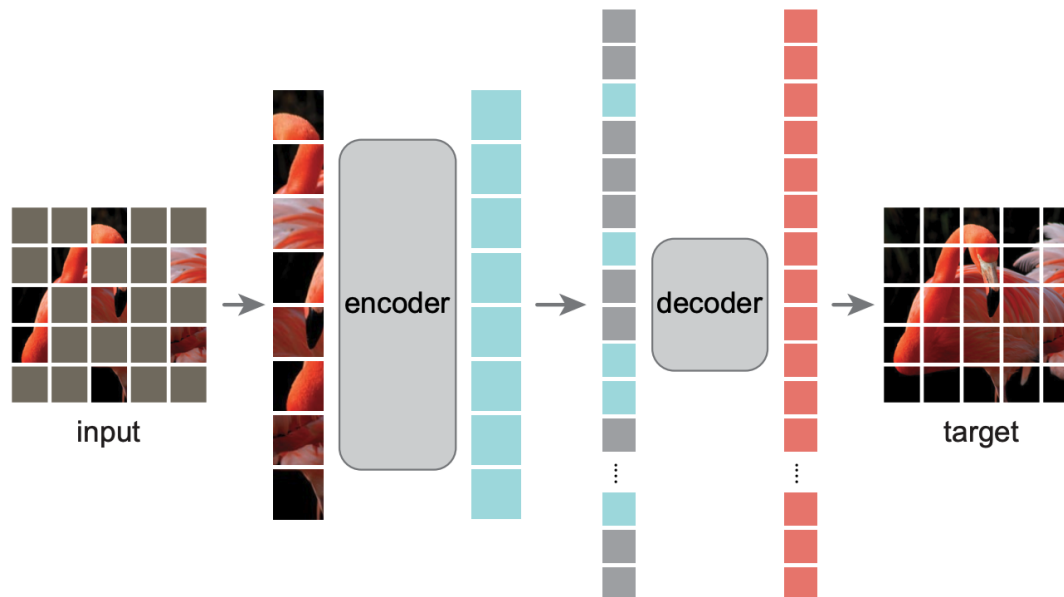
- Minimizes/maximizes the embedding distance of positive/negative pairs
- One of the first self-supervised methods for computer vision



Source: Barlow twins: Self-supervised learning via redundancy reduction

# Masked Image Methods

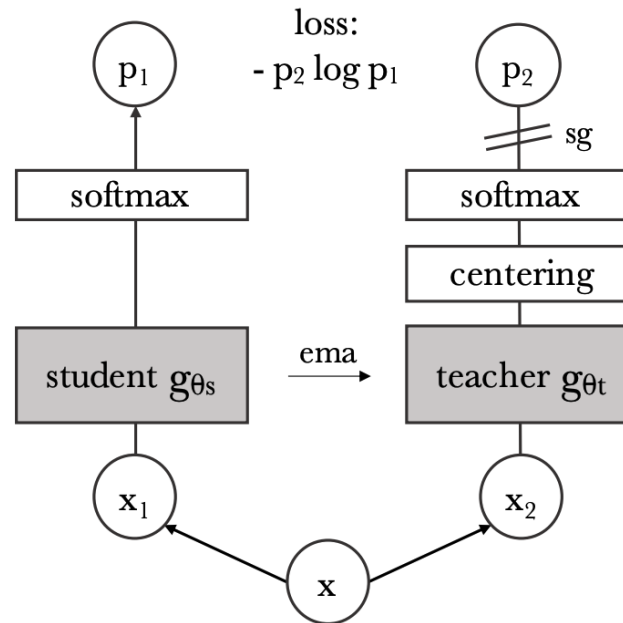
- Predicts the masked patches
- Motivated by the success of masking in LLM



Source: Masked autoencoders are scalable vision learners

# Self-Distillation Methods

- Uses an exponential moving average of the network weights as supervision
- Doesn't require negative samples

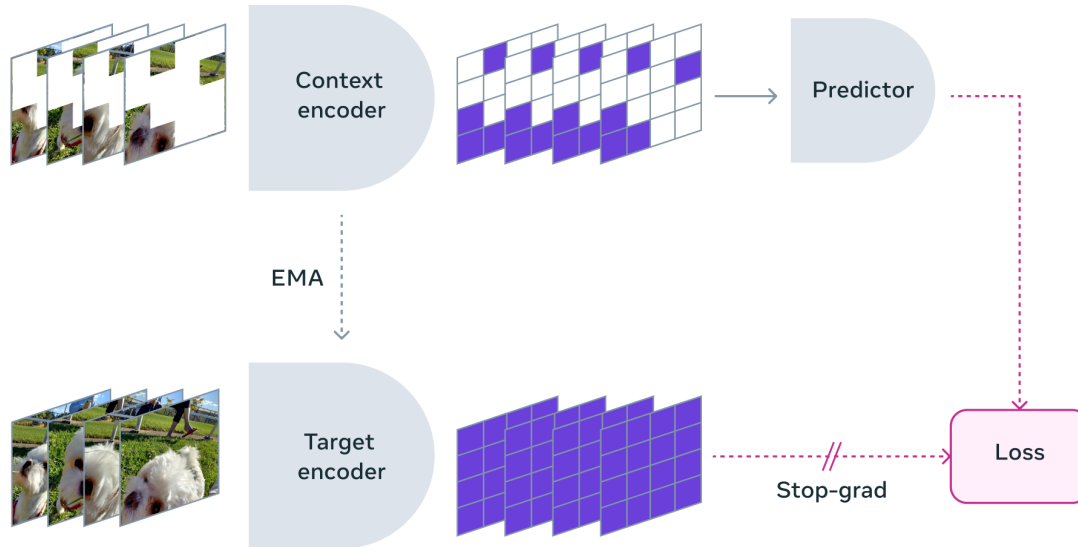


Source: Emerging properties in self-supervised vision transformers



# Video Methods

- Uses the temporal consistency in videos



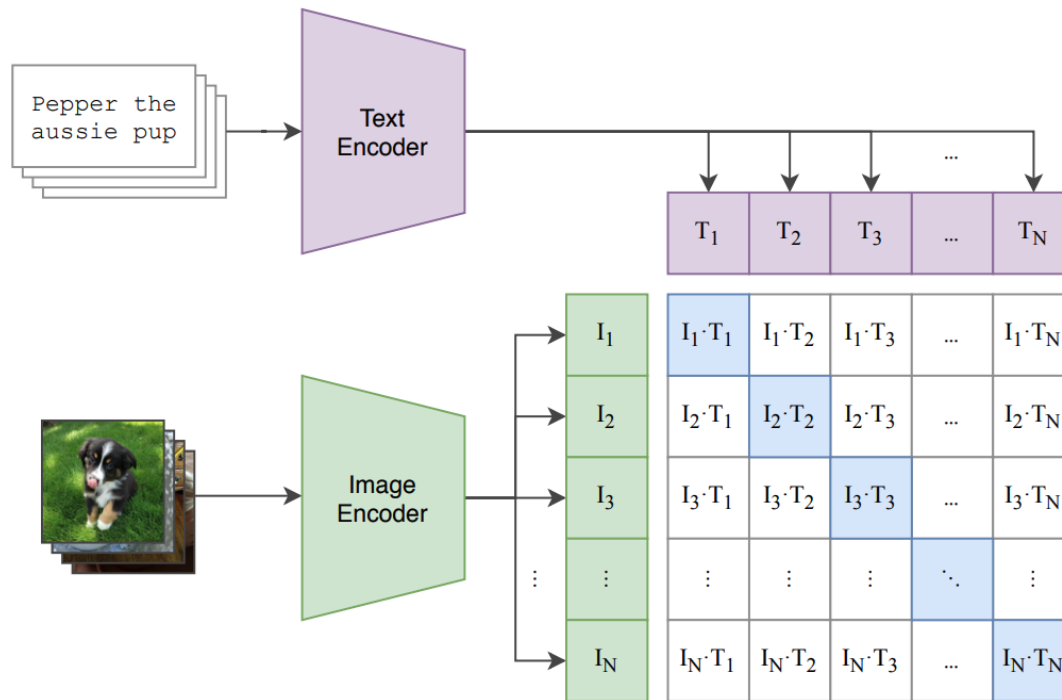
Source: V-JEPA: Latent video prediction for visual representation learning

# Outline

- General Information
  - About the Seminar
  - Registration
- Possible Topics
  - Self-Supervised Representation Learning in Computer Vision
  - Multi-Modal 2D Foundation Models
  - 3D Foundation Models
- Questions

# Visual Language Models

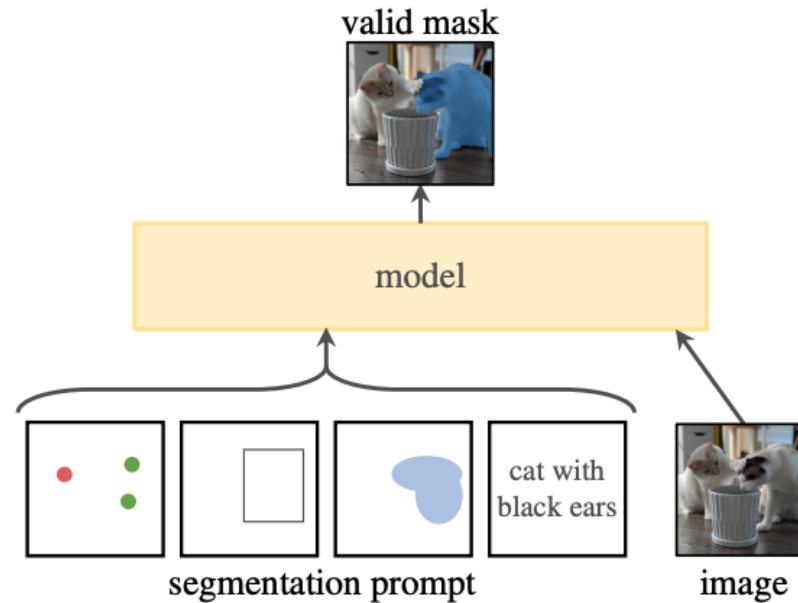
- Learns a shared embedding space of visual and language models



Source: Learning transferable visual models from natural language supervision

# Segmentation Models

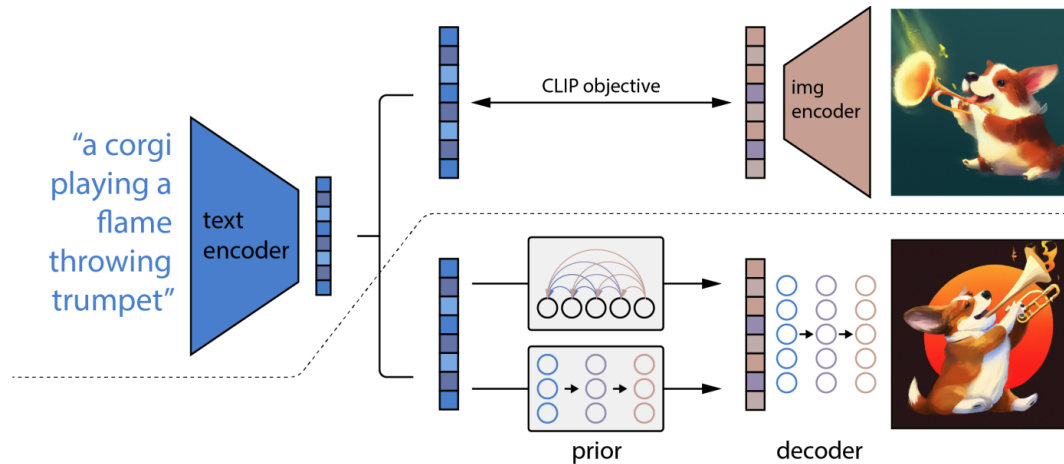
- Uses different user inputs to segment the object of interest
- Can be extended to work with language inputs



Source: Segment anything

# Generative Models

- Produces realistic images from text prompts
- Often uses diffusion models



Source: Hierarchical text-conditional image generation with clip latents

# Outline

- General Information
  - About the Seminar
  - Registration
- Possible Topics
  - Self-Supervised Representation Learning in Computer Vision
  - Multi-Modal 2D Foundation Models
  - 3D Foundation Models
- Questions

# 3D Foundation Models

- Produces realistic 3D models from text prompts



Source: Marigold: Repurposing diffusion-based image generators for monocular depth estimation

# Questions?

## Reminder:

- Web page: <https://cvg.cit.tum.de/teaching/ws2024/fmcv>
- Password: fmcv24neu
- Contact: [fmcv-ws24@vision.in.tum.de](mailto:fmcv-ws24@vision.in.tum.de)