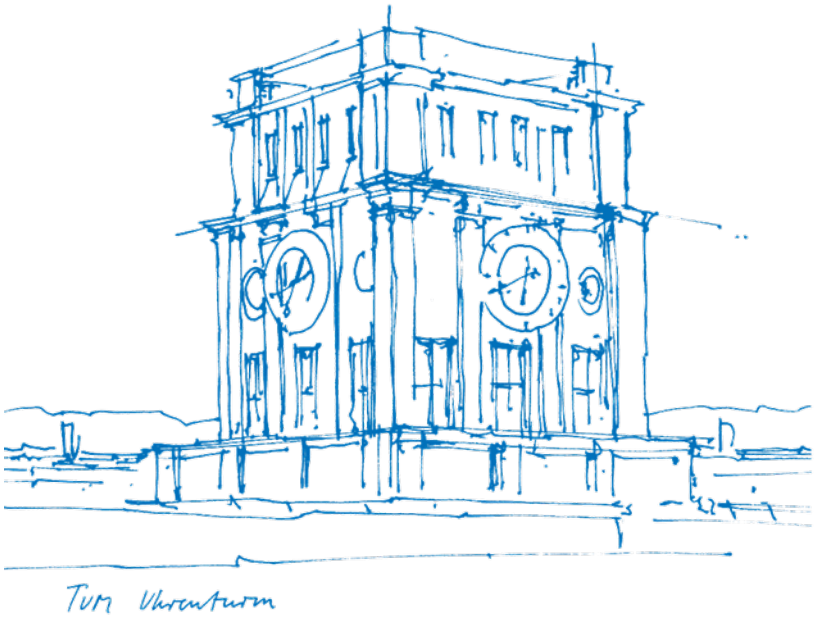
Seminar: Foundational Models for 2D and 3D Computer Vision

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



How can I access these slides?

- Option 1 (preferred): seminar web page
  - `https://cvg.cit.tum.de/teaching/ws2023/fmcv`
  - Password for the materials page: `ws23-fmcv`
  - Material page will go online after this pre-meeting
- Option 2: contact organizers
  - `fmcv-ws23@cvpr.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
  - Approximately 8 sessions (TBA)
  - **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 $\rightarrow$ 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 LaTeX template provided on web page
  - Length: 3-4 pages
  - Send final report as pdf by email to `fmcv-ws23@cvpr.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 25th, 2023

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**

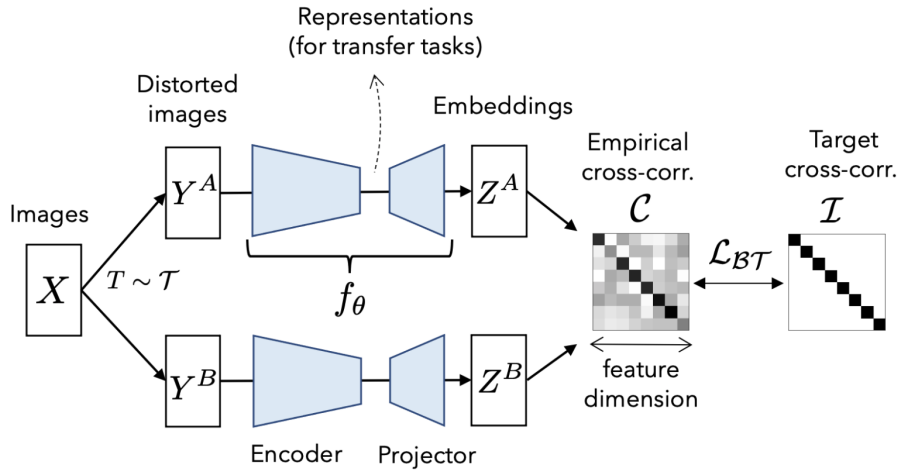
    emails containing all required information will be considered
  - Among students meeting the formal criteria, selection will be random
  - Note that if you have not taken any related course, you must be willing to invest a lot of work to learn the required basics
  - You will get notified by the matching system about the decision (28.07.2023)
- 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

Chen et al. *A simple framework for contrastive learning of visual representations*
Zbontar et al. *Barlow twins: Self-supervised learning via redundancy reduction*
He et al. *Momentum contrast for unsupervised visual representation learning*



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

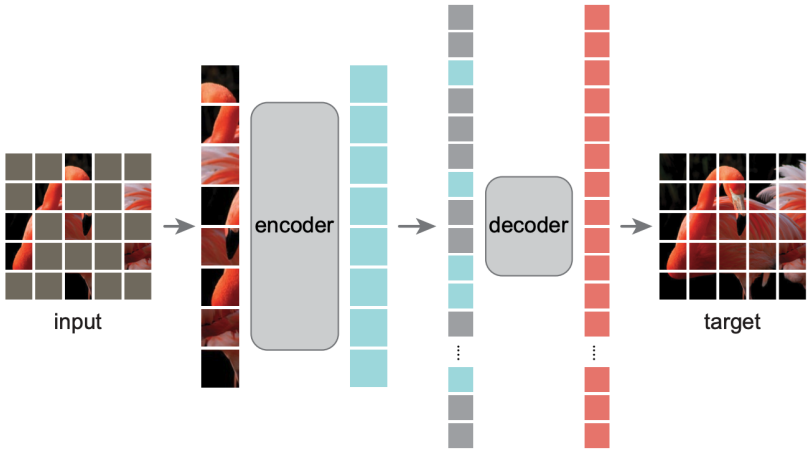
- Minimizes/maximizes the embedding distance of positive/negative pairs
- One of the first self-supervised methods for computer vision

Masked Image Methods

He et al. *Masked autoencoders are scalable vision learners*

Feichtenhofer et al. *Masked autoencoders as spatiotemporal learners*

Bao et al. *Beit: Bert pre-training of image transformers*



Source: Masked autoencoders are scalable vision learners

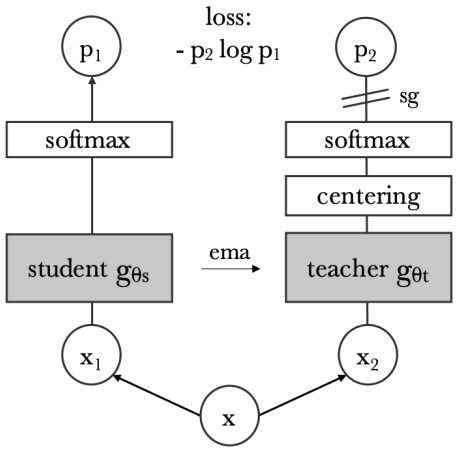
- Predicts the masked patches
- Motivated by the success of masking in LLM

Self-Distillation Methods

Grill et al. *Bootstrap your own latent-a new approach to self-supervised learning*

Caron et al. *Emerging properties in self-supervised vision transformers*

Oquab et al. *DINOv2: Learning Robust Visual Features without Supervision*



Source: Emerging properties in self-supervised vision transformers

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

Outline

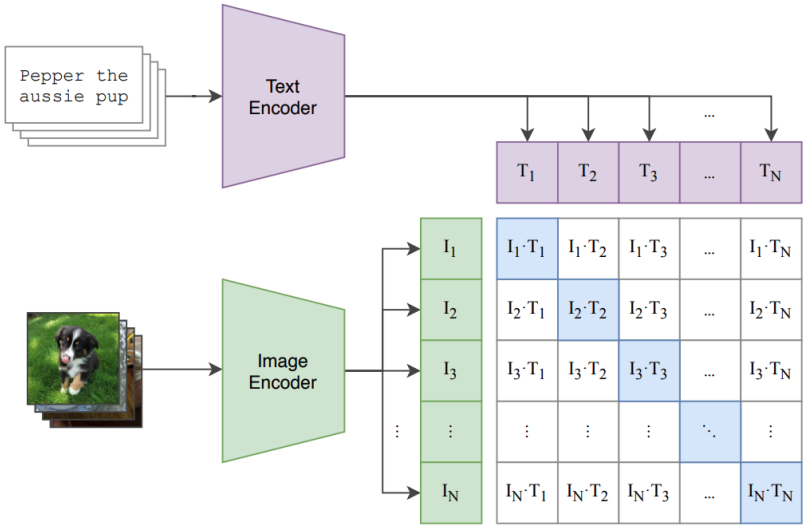
- General Information
  - About the Seminar
  - Registration
- Possible Topics
  - Self-Supervised Representation Learning in Computer Vision
  - Multi-Modal 2D Foundation Models
  - 3D Foundation Models
- Questions

Visual Language Models

Radford et al. *Learning transferable visual models from natural language supervision*

Yuan et al. *Florence: A new foundation model for computer vision*

Wang et al. *Image as a foreign language: Beit pretraining for all vision and vision-language tasks*



Source: Learning transferable visual models from natural language supervision

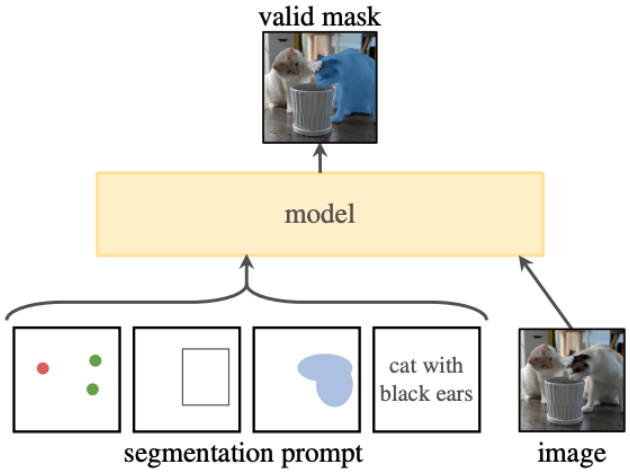
- Learns a shared embedding space of visual and language models

Segmentation Models

Kirillov et al. *Segment anything*

Zou et al. *Segment everything everywhere all at once*

Zhang et al. *Recognize Anything: A Strong Image Tagging Model*



Source: Segment anything

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

Generative Models

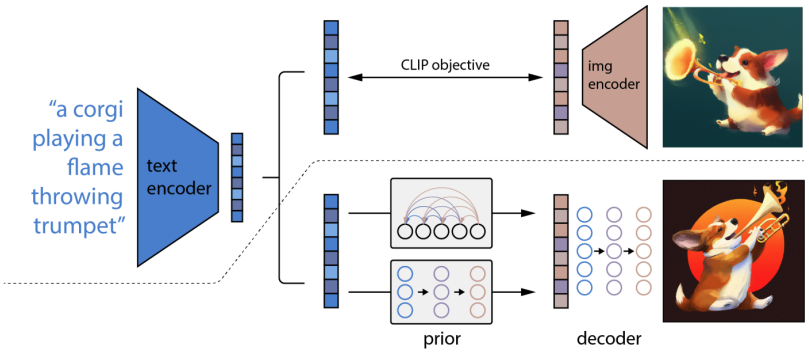
Ramesh et al. *Zero-shot text-to-image generation*

Nichol et al. *GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models*

Ramesh et al. *Hierarchical text-conditional image generation with clip latents*

Rombach et al. *High-resolution image synthesis with latent diffusion models*

Yu et al. *Scaling Autoregressive Models for Content-Rich Text-to-Image Generation*



Source: Hierarchical text-conditional image generation with clip latents

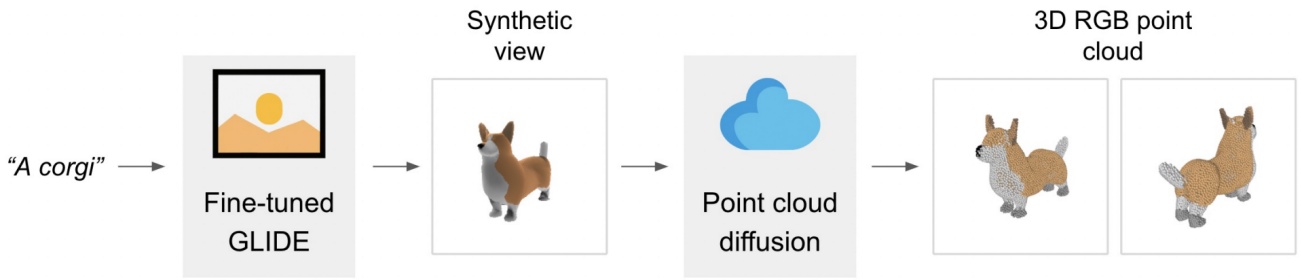
- Produces realistic images from text prompts

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 Generative Models

Poole et al. *DreamFusion: Text-to-3D using 2D Diffusion*
Nichol et al. *Point-E: A System for Generating 3D Point Clouds from Complex Prompts*
Liu et al. *MeshDiffusion: Score-based Generative 3D Mesh Modeling*
Singer et al. *Text-To-4D Dynamic Scene Generation*



Source: Point-E: A System for Generating 3D Point Clouds from Complex Prompts

- Produces realistic 3D models from text prompts

Questions?

Reminder:

- Web page: `https://cvg.cit.tum.de/teaching/ws2023/fmcv`
- Password: `ws23-fmcv`
- Contact: `fmcv-ws23@cvpr.in.tum.de`