1	**Visual Data Representation and Quality Assessment Metrics** Visual data comprises many different data types such as images, videos, point clouds, 360°-content, multispectral, and other representations. This topic shall present an overview over different visual data representations among with their mathematical descriptions. Also, a statistical view on visual data should be included, highlighting the manifold hypothesis and differences between natural and synthetic data. Furthermore, the key concepts of coding and quality assessment including frequency transformations, rate-distortion theory, and objective quality evaluation metrics shall be presented for a selection of data representations.
2	**Deep Learning: Basics** Deep Learning algorithms form the basis of most state-of-the-art computer vision systems. Hence, a basic understanding of the theory behind deep neural networks is essential for anyone working in this field. This topic will introduce the basics of neural networks by first explaining neurons, activations, and essential layers such as convolutional layers, pooling layers, and fully connected layers. The presentation will then cover how these networks can be trained using gradient-descent algorithms with the help of backpropagation. These foundational concepts are crucial for building and understanding advanced deep learning models.
3	**Deep Learning: Training and Risks** Effective training of deep neural networks is essential for achieving high performance in various applications. This topic shall introduce the fundamental principles and methodologies for training neural networks, focusing on different optimization algorithms such as Stochastic Gradient Descent (SGD), momentum, Adagrad and Adam, and explain how these algorithms build upon each other to improve training efficiency and convergence. Additionally, essential training techniques, including batch normalization, dropout, and learning rate schedules shall be covered, and common risks associated with training and deploying neural networks, such as class imbalance or adversarial attacks shall be addressed briefly.
4	**Self-Supervised and Contrastive Learning** Self-supervised learning enables deep neural networks to learn useful representations from unlabeled data by constructing supervisory signals directly from the data itself. Instead of relying on manually annotated labels, these methods train models through pretext tasks, data augmentations, masked prediction, or consistency objectives. Contrastive learning is an important family of self-supervised methods, where models learn to bring similar samples closer in representation space while pushing dissimilar samples apart. Representative approaches include SimCLR, Contrastive Predictive Coding, MoCo, BYOL, DINO, and masked autoencoder-based methods. These techniques have been widely used in image and video understanding, object detection, segmentation, denoising, super-resolution, and representation learning. Their main advantages include reduced annotation costs, improved transferability, and stronger robustness in downstream tasks. This topic shall explain the difference between supervised, self-supervised, and contrastive learning, introduce one or two representative methods such as SimCLR or DINO, and discuss how learned representations can be transferred to downstream vision tasks.

5	**Transfer Learning** Transfer learning is a central technique in deep learning where knowledge learned from one task, dataset, or domain is reused to improve performance on another related task. Instead of training a model from scratch, practitioners often start from a pretrained model and fine-tune it for a specific downstream task. This can reduce training time, lower data requirements, and improve accuracy, especially when labeled data is limited. In computer vision and multimedia, transfer learning is widely used for image classification, object detection, segmentation, video understanding, and image generation. Modern large-scale models further extend this idea through few-shot and zero-shot learning, where models generalize to new tasks with very few or no task-specific examples. Parameter-efficient fine-tuning methods, such as adapters, prompt tuning, and LoRA, adapt large models by updating only a small number of parameters. This topic shall explain the difference between training from scratch, feature extraction, full fine-tuning, and parameter-efficient fine-tuning. Furthermore, representative examples such as ImageNet-pretrained CNNs, CLIP-style zero-shot transfer, or LoRA-based adaptation, shall be discussed and the benefits and limitations of transfer learning in vision tasks shall be compared.
6	**Reinforcement Learning** Reinforcement learning (RL) is a learning paradigm in which an agent learns to make decisions by interacting with an environment and receiving rewards. It is commonly formulated using Markov Decision Processes (MDPs), while Partially Observable MDPs (POMDPs) are used when the agent has incomplete observations of the environment. Core RL methods include value-based approaches such as Q-learning and Deep Q-Networks (DQN), as well as policy-based approaches such as REINFORCE and actor-critic methods. More advanced algorithms, including Proximal Policy Optimization (PPO) and Advantage Actor-Critic (A2C), improve the stability and efficiency of policy learning. In computer vision and robotics, RL is used for visual navigation, robotic manipulation, active object recognition, interaction, video understanding, and decision-making under uncertainty. This topic shall explain the basic RL framework, including states, actions, rewards, policies, and value functions. It shall introduce representative algorithms such as Q-learning, DQN, REINFORCE, PPO, or A2C, and discuss how visual observations can be used for robot navigation, manipulation, and interaction tasks.
7	**Physics-Informed Neural Networks** Physics-informed neural networks (PINNs) integrate physical laws into the training process of neural networks. Instead of learning only from data, PINNs are trained with additional constraints derived from governing equations, such as conservation laws or partial differential equations. In visual computing, this idea is especially relevant for generating or predicting physically plausible visual data, such as smoke, water, fire, cloth, or deformable objects. For example, fluid motion can be constrained by equations from fluid dynamics, helping the model produce results that are not only visually realistic but also physically consistent. PINNs can reduce the need for large labeled datasets and improve generalization in scientific simulation, computer graphics, robotics, and video prediction. However, challenges remain in optimization stability, computational cost, and scaling to complex real-world scenes. This topic shall explain the basic idea of physics-informed learning and how physical constraints can be added to neural network training. It shall discuss examples such as smoke or water simulation, introduce the role of governing equations like fluid dynamics constraints, and compare PINNs with purely data-driven methods in terms of realism, physical consistency, and computational cost.

8	**Gaussian Splatting** Gaussian Splatting is an explicit scene representation and rendering technique that models a 3D scene as a collection of learnable Gaussian primitives. Each Gaussian typically contains position, scale, rotation, opacity, and appearance information, allowing scenes to be rendered efficiently through differentiable splatting instead of dense volumetric sampling. 3D Gaussian Splatting has become popular for real-time novel-view synthesis, 3D reconstruction, and neural rendering, offering a practical alternative to NeRF-style implicit representations. The idea can also be extended to dynamic scenes and videos through 4D Gaussian representations, where time-varying geometry and appearance are modeled jointly. Since Gaussian scenes may contain millions of primitives, storage and transmission become important challenges, making compression of Gaussian parameters, point-cloud-like structures, and rendered images/videos an active research direction. This topic shall cover how Gaussian Splatting represents scenes differently from meshes, point clouds, and NeRFs. It shall discuss the role of Gaussian parameters, differentiable rendering, and real-time novel-view synthesis. Possible extensions include dynamic/video Gaussian Splatting, point-cloud-based initialization, and compression methods for reducing storage, bandwidth, and rendering cost.
9	**Implicit Neural Representations** Implicit neural representations (INRs) parameterize signals as continuous functions using neural networks. Instead of storing visual data as discrete pixels, voxels, meshes, or frames, INRs map input coordinates directly to signal values such as color, density, occupancy, or intensity. This makes them suitable for representing images, audio, videos, 3D shapes, and radiance fields in a resolution-independent manner. A key advantage of INRs is that they can encode complex signals compactly while incorporating useful priors through the network architecture, activation functions, and positional encodings. Representative methods include SIREN, which uses periodic activations to model high-frequency details, NeRF for novel-view synthesis, NeRV for video representation, and Cool-Chic for image compression. INRs are especially relevant for neural rendering, signal compression, 3D reconstruction, and generative visual representation. This topic shall explain how INRs differ from traditional discrete representations such as images, voxels, meshes, and point clouds. It shall introduce representative methods such as SIREN, NeRF, NeRV, or Cool-Chic, and discuss the advantages and limitations of coordinate-based neural representations in rendering, compression, and 3D scene modeling.
10	**Object Detection and Classification in the Pixel-domain** Object detection and classification are foundational tasks in computer vision, enabling systems to identify and categorize objects within images and videos. This topic shall introduce the fundamental concepts of object detection and classification, cover key frameworks such as YOLO and Faster R-CNN, and explain their architectures, strengths, and applications. The topic shall also provide a brief overview of common object tracking methodologies such as SORT (Simple Online and Realtime Tracking), Deep SORT, or tracking-by-detection.

11	**Image Processing with Latent Representations** Latent-space image processing refers to methods that transform images into compact feature representations before performing downstream tasks. Instead of directly operating on raw pixels, neural networks learn latent representations that capture semantic, structural, or textural information from the input image. These representations can be obtained from autoencoders, variational autoencoders, convolutional neural networks (CNNs), vision transformers, or self-supervised models. Working in latent space can make image processing more efficient and robust, because irrelevant pixel-level variations may be suppressed while useful high-level features are preserved. Applications include texture prediction, image classification, face detection, image retrieval, image editing, denoising, super-resolution, and generative modeling. This topic shall also cover how the quality of the learned latent space affects task performance, interpretability, compression, and generalization. Furthermore, a comparison to pixel-space processing for tasks such as texture prediction, image classification, and face detection shall be presented. This topic shall explain what a latent representation is and why it is useful for image processing. It shall discuss how models such as CNNs, autoencoders, VAEs, or vision transformers encode images into latent features, and compare pixel-space processing with latent-space processing for tasks such as texture prediction, image classification, and face detection.
12	**Depth Perception and Estimation** Depth perception is essential for autonomous systems to understand the 3D structure of their environment and to position themselves relative to surrounding objects. It enables applications such as autonomous driving, robot navigation, obstacle avoidance, augmented reality, 3D reconstruction, and scene understanding. Depth can be obtained using dedicated hardware sensors such as LiDAR, radar, time-of-flight cameras, or structured-light depth cameras. Another important direction is image-based depth estimation, where depth is inferred from visual information. Multi-camera systems, such as stereo cameras, estimate depth by matching corresponding points across different views and using geometric triangulation. Monocular depth estimation aims to predict depth from a single RGB image, often relying on learned priors from deep neural networks. Students should also explore sensor fusion, where information from cameras and depth sensors is combined to improve robustness and accuracy in real-world systems. This topic shall compare hardware-based depth sensing, multi-view or stereo depth estimation, and single-image depth prediction. It shall explain the basic geometry of stereo vision, the role of camera calibration and correspondence matching, and how sensor fusion can combine RGB cameras, LiDAR, radar, or depth cameras for reliable perception in autonomous systems.
13	**Human Action Recognition** Human action recognition aims to identify and understand human activities from videos, images, or sensor data. Unlike static image recognition, this task requires modeling both spatial appearance and temporal motion information. Students should explore common video-based approaches, including two-stream networks that combine RGB frames and optical flow, 3D convolutional networks for spatiotemporal feature learning, and recurrent or Transformer-based models for capturing long-range temporal dependencies. Action localization is also an important extension: temporal action localization detects the start and end times of actions in videos, while spatial action localization estimates where the action occurs in each frame. Skeleton-based methods using graph convolutional networks represent the human body as a graph and model the relations between joints. Current challenges include untrimmed videos, occlusion, viewpoint variation, ambiguous actions, limited annotations, and few-shot generalization. This topic shall explain the difference between action recognition, temporal action localization, and spatial action localization. It shall introduce representative model families such as two-stream networks, 3D CNNs, LSTMs, Transformers, and graph convolutional networks, and discuss how different methods capture motion, appearance, and human pose information.

14	**Optimal Transport Theory for Generative Models** Optimal transport theory builds the basis for any generative model where samples are drawn from a probability distribution or translated to a match some desired statistical or visual properties. This topic shall cover the mathematical basics of optimal transport theory and distance metrics between probability distributions like Kullback-Leibler divergence or Wasserstein distance. Furthermore, the connection of optimal transport theory to a selection of generative models shall be covered.
15	**Deep Generative Models** Generative models "invent" new data from a desired distribution, optionally conditioned by some constraints like an image template or a text prompt. This topic shall serve as an overview over different approaches to deep generation of visual data and their corresponding training procedures. In the initial phase of deep image generation, generative adversarial networks (GANs) were proposed that let a generator, that creates new data, and a discriminator, that labels its input as real or fake, compete against each other during training until an equilibrium is reached. Recent approaches build upon diffusion models that start from noise and iteratively converge towards a generated sample. Typical applications of deep generative models are image-to-image translation or text-to-image.
16	**Deep Fakes** Deep fakes have received a lot of attention through social media due to the integrity risks, especially for politicians or other highly influential persons. This topic shall introduce recent generative models that are able to create deep fakes as well as cover security risks and guidelines for the use of such models. Furthermore, deep fake detection methods should also be covered.
17	**Inverse Problems** Inverse problems aim on reverting any image degradation process like noise, blur, or loss of color information. Since lost information has to be restored, deep generative models perform best for this task. This topic shall introduce some inverse problems like denoising, deblurring, and image restoration and present models that offer a solution for the corresponding problem. Since medical images often suffer from high noise-levels and other distortions, they can serve as a common application of inverse problems.
18	**Learning-based Image Compression** Learned image compression applies deep learning techniques to compress images more efficiently than traditional methods. This approach typically uses neural network architectures like autoencoders or transformers to learn optimal representations of image data. Key components include entropy models for latent representations and encoding/decoding networks. Learned compression methods have shown superior rate-distortion performance compared to classical standards like JPEG or BPG, often preserving more details at lower bitrates. Recent advancements incorporate attention mechanisms, context modeling, and hybrid CNN-Transformer architectures to further improve compression efficiency.
19	**Learning-based Video Compression** Learned video compression extends learned image compression techniques by exploiting temporal correlations. This topic shall cover the basics of optical flow-based motion estimation and motion compensation, essential for effectively compressing video by predicting and encoding motion between frames. Furthermore, the concepts of residual coding and conditional coding shall be introduced among an overview over the most influential networks in the field, including Deep Video Compression (DVC), Deep Contextual Video Compression (DCVC), as well as its extensions DCVC-TCM, -HEM, -DC, -FM, and the state of the art DCVC-RT.

20	**Semantic and Generative Coding** Semantic and generative coding explores new paradigms for image and video compression beyond traditional pixel-level reconstruction. Instead of only minimizing distortion between original and decoded pixels, semantic coding aims to preserve task-relevant or meaning-level information, such as objects, actions, scene structure, or features used by machine vision systems. Generative coding further uses powerful generative models, such as GANs, VAEs, diffusion models, or large multimodal models, to reconstruct visually plausible content from compact latent codes, semantic descriptions, or sparse visual information. These methods are especially relevant for ultra-low-bitrate compression, visual communication, video coding for machines, and human-machine collaborative perception. Students may discuss how semantic representations, learned latents, text prompts, motion cues, and diffusion-based decoders can reduce bitrate while maintaining perceptual quality or downstream task performance. This topic shall explain the difference between traditional signal-level coding, semantic coding, and generative coding. It shall discuss representative ideas such as learned latent compression, video coding for machines, feature-based compression, and diffusion- or GAN-based reconstruction. It shall also briefly compare evaluation criteria, including PSNR/MS-SSIM, perceptual quality, FID/LPIPS-style metrics, and downstream task accuracy for detection, segmentation, or recognition.
21	**Processing and Coding of 360°-Videos** In contrast to standard videos with limited field of view, 360°- or omnidirectional videos capture light from all directions. The underlying spherical geometry of 360°-videos makes it impossible to process them with well-known tools for 2D videos without introducing any geometrical distortion. This topic shall introduce the mathematical description of 360°-videos and common projection formats like equirectangular- (ERP), cubemap- (CMP), equiangular cubemap- (EAC), generalized cubemap- (GCP), and equatorial cylindrical projection (ECP) and showcase some distortions that make the processing and coding of spherical content challenging when the sphere is projected onto a planar surface. Furthermore, a selection of processing and coding algorithms that are adapted to 360°-images or videos shall be presented.
22	**Perception-based Quality Evaluation** Objective quality evaluation metrics aim to quantify the reconstruction quality of a video to match subjective human perception. However, it is an open research topic how to most accurately match the human perception. This topic shall cover different approaches beyond pixel fidelity, captured by PSNR, towards accurately matching human perception, such as BRISQUE and NIQE for no-reference metrics, and MS-SSIM, LPIPS, and VMAF for full-reference metrics.

23	**Processing and Compression of Astronomical Data** Astronomical data processing focuses on extracting scientifically meaningful information from large-scale observations collected by telescopes, satellites, and sensor arrays. Unlike ordinary natural images, astronomical data often has high dynamic range, low signal-to-noise ratio, sparse structures, and strong noise or artifact patterns caused by sensors, atmosphere, or acquisition conditions. This topic shall explore the basic processing pipeline, including calibration, denoising, background subtraction, source detection, image registration, and data visualization. Compression is another central challenge, since modern sky surveys and space missions generate massive amounts of image, spectral, time-series, and radio data. Both lossless and lossy compression methods are relevant, but scientific fidelity must be carefully preserved. Representative topics include FITS image compression, Rice/Hcompress-style classical algorithms, learned image compression, and task-aware compression for astronomical analysis. This topic shall also explain why astronomical data is different from standard visual data, especially in terms of noise, dynamic range, sparsity, and scientific accuracy. It shall introduce typical processing steps such as calibration, denoising, source detection, and image alignment, and then discuss compression methods for reducing storage and transmission costs while preserving information needed for scientific analysis.