Offline Session M1

Paper id	Paper title
mmfp1731	U2Net: A General Framework with Spatial-Spectral-Integrated Double U-Net for Image Fusion
mmfp1796	Practical Edge Detection via Robust Asynchronous Learning
mmfp1801	Generalizing Face Forgery Detection via Uncertainty Learning
mmfp1807	M3R: Masked Token Mixup and Cross-Modal Reconstruction for Zero-Shot Learning
mmfp1812	Regress Before Construct: Regress Autoencoder for Point Cloud Self-supervised Learning
mmfp1817	Transferring CLIP's Knowledge into Zero-Shot Point Cloud Semantic Segmentation
mmfp1876	PVG: Progressive Vision Graph for Vision Recognition
mmfp1880	Causal Intervention for Sparse-View Gait Recognition
mmfp1912	Graph-Based Video-Language Learning with Multi-Grained Audio-Visual Alignment
mmfp1935	M2ATS: A Real-world Multimodal Air Traffic Situation Benchmark Dataset and Beyond
mmfp1968	Event-Enhanced Multi-Modal Spiking Neural Network for Dynamic Obstacle Avoidance
mmfp2082	Occluded Skeleton-Based Human Action Recognition with Dual Inhibition Training
mmfp2129	Text-Only Training for Visual Storytelling
mmfp2155	HELIOS: Hyper-Relational Schema Modeling from Knowledge Graphs
mmfp2162	A Unified Query-based Paradigm for Camouflaged Instance Segmentation
mmfp2178	Bidomain Modeling Paradigm for Pansharpening
mmfp2183	Multi-view Graph Clustering via Efficient Global-Local Spectral Embedding Fusion
mmfp2214	Uni-paint: A Unified Framework for Multimodal Image Inpainting with Pretrained Diffusion Model
mmfp2296	Exploring Dual Representations in Large-Scale Point Clouds: A Simple Weakly Supervised Semantic Segmentation Framework
mmfp2329	Exploring Universal Principles for Graph Contrastive Learning: A Statistical Perspective
mmfp2349	Preserving Local and Global Information: An Effective Metric-based Subspace Clustering
mmfp2361	Advancing Video Question Answering with a Multi-modal and Multi-layer Question Enhancement Network
mmfp2378	Weakly-Supervised Text Instance Segmentation
mmfp2405	Chaos to Order: A Label Propagation Perspective on Source-Free Domain Adaptation
mmfp2450	Mutual-guided Dynamic Network for Image Fusion
mmfp2461	SA-GDA: Spectral Augmentation for Graph Domain Adaptation
mmfp2526	GraMMaR: Ground-aware Motion Model for 3D Human Motion Reconstruction
mmfp2534	Learning Generalized Representations for Open Set Temporal Action Localization
mmfp2594	Boosting Few-shot 3D Point Cloud Segmentation via Query-Guided Enhancement
mmfp2670	Mutual Information-driven Triple Interaction Network for Efficient Image Dehazing
mmfp2752	Fine-Grained Multimodal Named Entity Recognition and Grounding with a Generative Framework
mmfp2788	Dropping Pathways Towards Deep Multi-View Graph Subspace Clustering Networks
mmfp2795	Uni-Dual: A Generic Unified Dual-Task Medical Self-Supervised Learning Framework
mmfp2839	Text-to-Audio Generation using Instruction Guided Latent Diffusion Model
mmfp2849	Improving Federated Person Re-Identification through Feature-Aware Proximity and Aggregation
mmfp2871	P2I-NET: Mapping Camera Pose to Image via Adversarial Learning for New View Synthesis in Real Indoor Environments
mmfp2888	FSR-Net: Deep Fourier Network for Shadow Removal
mmfp2953	Single Domain Generalization via Unsupervised Diversity Probe
mmfp2980	A Figure Skating Jumping Dataset for Replay-Guided Action Quality Assessment
mmfp3001	Train One, Generalize to All: Generalizable Semantic Segmentation from Single-Scene to All Adverse Scenes
mmfp3046	PMVC: Data Augmentation-Based Prosody Modeling for Expressive Voice Conversion
mmfp3107	Towards Better Multi-modal Keyphrase Generation via Visual Entity Enhancement and Multi-granularity Image Noise Filtering
mmfp3143	PDE-based Progressive Prediction Framework for Attribute Compression of 3D Point Clouds
mmfp3146	Uncertainty-Guided End-to-End Audio-Visual Speaker Diarization for Far-field Recordings
mmfp3199	CTCP: Cross Transformer and CNN for Pansharpening
mmfp3214	Video Frame Interpolation with Flow Transformer
mmfp3317	Event-Diffusion: Event-Based Image Reconstruction and Restoration with Diffusion Models
mmfp3323	Towards Balanced Active Learning for Multimodal Classification
mmfp3331	Saliency Prototype for RGB-D and RGB-T Salient Object Detection
mmfp3335	MCG-MNER: A Multi-Granularity Cross-Modality Generative Framework for Multimodal NER with Instruction
mmfp3429	Distribution Consistency based Fast Anchor Imputation for Incomplete Multi-view Clustering
mmfp3441	POV: Prompt-Oriented View-agnostic Learning for Egocentric Hand-Object Interaction in the Multi-view World
mmfp3456	Debunking Free Fusion Myth: Online Multi-view Anomaly Detection with Disentangled Product-of-Experts Modeling
mmfp3494	Multi-View Representation Learning via View-Aware Modulation
mmfp3591	DeepSVC: Deep Scalable Video Coding for Both Machine and Human Vision
mmfp3600	Induction Network: Audio-Visual Modality Gap-Bridging for Self-Supervised Sound Source Localization
mmfp3621	BLAT: Bootstrapping Language-Audio Pre-training based on AudioSet Tag-guided Synthetic Data
mmfp3680	Breaking the Barrier Between Pre-training and Fine-tuning: A Hybrid Prompting Model for Knowledge-Based VQA
mmfp3790	Dynamic Compositional Graph Convolutional Network for Efficient Composite Human Motion Prediction
mmfp3827	OccluBEV: Occlusion Aware Spatiotemporal Modeling for Multi-view 3D Object Detection
mmfp3837	ELFIC: A Learning-based Flexible Image Codec with Rate-Distortion-Complexity Optimization
mmfp3877	Localized and Balanced Efficient Incomplete Multi-view Clustering
mmfp3888	Variance-Aware Bi-Attention Expression Transformer for Open-Set Facial Expression Recognition in the Wild
mmfp3948	Universal Domain Adaptive Network Embedding for Node Classification
mmfp3971	Skeletal Spatial-Temporal Semantics Guided Homogeneous-Heterogeneous Multimodal Network for Action Recognition
mmfp4026	Multi-modal Social Bot Detection: Learning Homophilic and Heterophilic Connections Adaptively

mmfp2061	Exploring the Knowledge Transferred by Response-Based Teacher-Student Distillation
mmfp2080	StylePrompter: All Styles Need Is Attention
mmfp2087	TIRDet: Mono-Modality Thermal InfraRed Object Detection Based on Prior Thermal-To-Visible Translation
mmfp2099	Shifted GCN-GAT and Cumulative-Transformer based Social Relation Recognition for Long Videos
mmfp2180	Selecting Learnable Training Samples is All DETRs Need in Crowded Pedestrian Detection

mmfp2242	ASTDF-Net: Attention-Based Spatial-Temporal Dual-Stream Fusion Network for EEG-Based Emotion Recognition
mmfp2247	Beware of Overcorrection: Scene-induced Commonsense Graph for Scene Graph Generation
mmfp2250	Collaborative Learning of Diverse Experts for Source-free Universal Domain Adaptation
mmfp2287	M^{3}Net: Multi-view Encoding, Matching, and Fusion for Few-shot Fine-grained Action Recognition
mmfp2345	Consistency-aware Feature Learning for Hierarchical Fine-grained Visual Classification
mmfp2351	Lightweight Super-Resolution Head for Human Pose Estimation
mmfp2354	Multi-view Self-Expressive Subspace Clustering Network
mmfp2385	CropCap: Embedding Visual Cross-Partition Dependency for Image Captioning
mmfp2444	UniNeXt: Exploring A Unified Architecture for Vision Recognition
mmfp2596	Automatic Network Architecture Search for RGB-D Semantic Segmentation
mmfp2712	Exploring High-Correlation Source Domain Information for Multi-Source Domain Adaptation in Semantic Segmentation
mmfp2722	On Regularizing Multiple Clusterings for Ensemble Clustering by Graph Tensor Learning
mmfp2747	Cross-modal & Cross-domain Learning for Unsupervised LiDAR Semantic Segmentation
mmfp2780	Cross-Modal and Multi-Attribute Face Recognition: A Benchmark
mmfp2818	Quality-Aware RGBT Tracking via Supervised Reliability Learning and Weighted Residual Guidance
mmfp2824	DLFusion: Painting-Depth Augmenting-LiDAR for Multimodal Fusion 3D Object Detection
mmfp2853	Generating Explanations for Embodied Action Decision from Visual Observation
mmfp2900	RAHNet: Retrieval Augmented Hybrid Network for Long-tailed Graph Classification
mmfp2906	MVCIR-net: Multi-view Clustering Information Reinforcement Network
mmfp2944	Masked Text Modeling: A Self-Supervised Pre-training Method for Scene Text Detection
mmfp2972	Multi-Speed Global Contextual Subspace Matching for Few-Shot Action Recognition
mmfp2985	Perceiving Ambiguity and Semantics without Recognition: An Efficient and Effective Ambiguous Scene Text Detector
mmfp3020	Learning Causality-inspired Representation Consistency for Video Anomaly Detection
mmfp3055	Deep Image Harmonization in Dual Color Spaces
mmfp3142	HARP: Let Object Detector Undergo Hyperplasia to Counter Adversarial Patches
mmfp3144	Hierarchical Semantic Enhancement Network for Multimodal Fake News Detection
mmfp3157	Unifying Two-Stream Encoders with Transformers for Cross-Modal Retrieval
mmfp3163	Bio-Inspired Audiovisual Multi-Representation Integration via Self-Supervised Learning
mmfp3182	Multi-teacher Self-training for Semi-supervised Node Classification with Noisy Labels
mmfp3185	Multi-scale Spatial-Spectral Attention Guided Fusion Network for Pansharpening
mmfp3213	Scene-aware Human Pose Generation using Transformer
mmfp3229	AffectFAL: Federated Active Affective Computing with Non-IID data
mmfp3258	Unified Multi-modal Unsupervised Representation Learning for Skeleton-based Action Understanding
mmfp3280	Painterly Image Harmonization using Diffusion Model
mmfp3333	Intra- and Inter-Modal Curriculum for Multimodal Learning
mmfp3356	Cross-modal and Cross-medium Adversarial Attack for Audio
mmfp3442	SpeechTripleNet: End-to-End Disentangled Speech Representation Learning for Content, Timbre and Prosody
mmfp3487	Transformer-based Open-world Instance Segmentation with Cross-task Consistency Regularization
mmfp3532	Data-Efficient Masked Video Modeling for Self-supervised Action Recognition
mmfp3585	Read Ten Lines at One Glance: Line-Aware Semi-Autoregressive Transformer for Multi-Line Handwritten Mathematical Expression Recognition
mmfp3610	Curriculum-Listener: Consistency- and Complementarity-Aware Audio-Enhanced Temporal Sentence Grounding
mmfp3770	Your tone speaks louder than your face! Modality Order Infused Multi-modal Sarcasm Detection
mmfp3787	Cross-Illumination Video Anomaly Detection Benchmark
mmfp3810	Uncertainty-Aware Variate Decomposition for Self-supervised Blind Image Deblurring
mmfp3890	Multi-Scale Similarity Aggregation for Dynamic Metric Learning
mmfp3974	Little Strokes Fell Great Oaks: Boosting the Hierarchical Features for Multi-exposure Image Fusion
mmfp4062	DRIN: Dynamic Relation Interactive Network for Multimodal Entity Linking
mmfp4065	Think before You Leap: Content-Aware Low-Cost Edge-Assisted Video Semantic Segmentation

Offline Session M2

Paper id	Paper title
mmfp1358	pmBQA: Projection-based Blind Point Cloud Quality Assessment via Multimodal Learning
mmfp0024	A Simple Baseline for Open-World Tracking via Self-training
mmfp0080	Explicify Neural Implicit Fields for Efficient Dynamic Human Avatar Modeling via a Neural Explicit Surface
mmfp0095	Modal-aware Visual Prompting for Incomplete Multi-modal Brain Tumor Segmentation
mmfp0099	CUCL: Codebook for Unsupervised Continual Learning
mmfp0155	Disentangled Representation Learning with Causality for Unsupervised Domain Adaptation
mmfp0188	High-order Complementarity Induced Fast Multi-View Clustering with Enhanced Tensor Rank Minimization
mmfp0254	Reparo: QoE-Aware Live Video Streaming in Low Rate Networks by Intelligent Frame Recovery
mmfp0285	Mask to reconstruct: Cooperative Semantics Completion for Video-text Retrieval
mmfp0296	Motion-Decoupled Spiking Transformer for Audio-Visual Zero-Shot Learning
mmfp0358	Frequency Representation Integration for Camouflaged Object Detection
mmfp0373	Rethinking Voice-Face Correlation: A Geometry View
mmfp0380	Attentive Alignment Network for Multispectral Pedestrian Detection
mmfp0403	Robust Spectral Embedding Completion Based Incomplete Multi-view Clustering
mmfp0406	MEAformer: Multi-modal Entity Alignment Transformer for Meta Modality Hybrid
mmfp0424	Calibration-based Dual Prototypical Contrastive Learning Approach for Domain Generalization Semantic Segmentation
mmfp0428	Disentangling Multi-view Representations Beyond Inductive Bias
mmfp0433	Exploring Hyperspectral Histopathology Image Segmentation from A Deformable Perspective
mmfp0458	Alleviating Spatial Misalignment and Motion Interference for UAV-based Video Recognition
mmfp0504	CONVERT:Contrastive Graph Clustering with Reliable Augmentation
mmfp0507	Partitioned Saliency Ranking with Dense Pyramid Transformers
mmfp0522	Human-Object-Object Interaction: Towards Human-Centric Complex Interaction Detection
mmfp0657	LandmarkGait: Intrinsic Human Parsing for Gait Recognition
mmfp0658	Interpolation Normalization for Contrast Domain Generalization
mmfp0666	Triple-Granularity Contrastive Learning for Deep Multi-View Subspace Clustering
mmfp0715	TMac: Temporal Multi-Modal Graph Learning for Acoustic Event Classification
mmfp0796	VTLayout: A Multi-Modal Approach for Video Text Layout
mmfp0814	Mixture-of-Experts Learner for Single Long-Tailed Domain Generalization
mmfp0815	RefineTAD: Learning Proposal-free Refinement for Temporal Action Detection
mmfp0819	Visual Causal Scene Refinement for Video Question Answering
mmfp0846	FeaCo: Reaching Robust Feature-Level Consensus in Noisy Pose Conditions
mmfp0850	Hermes: Leveraging Implicit Inter-Frame Correlation for Bandwidth-Efficient Mobile Volumetric Video Streaming
mmfp0900	CgT-GAN: CLIP-guided Text GAN for Image Captioning
mmfp0921	Skeleton MixFormer: Multivariate Topology Representation for Skeleton-based Action Recognition
mmfp0965	ALA: Naturalness-aware Adversarial Lightness Attack
mmfp0972	Mamba: Bringing Multi-dimensional ABR to WebRTC
mmfp1035	PAIF: Perception-Aware Infrared-Visible Image Fusion for Attack-Tolerant Semantic Segmentation
mmfp1090	Parameter-Efficient Transfer Learning for Audio-Visual-Language Tasks
mmfp1115	Foreground/Background-Masked Interaction Learning for Spatio-temporal Action Detection
mmfp1137	DealMVC: Dual Contrastive Calibration for Multi-view Clustering
mmfp1171	Informative Classes Matter: Towards Unsupervised Domain Adaptive Nighttime Semantic Segmentation
mmfp1211	CALM: An Enhanced Encoding and Confidence Evaluating Framework for Trustworthy Multi-view Learning
mmfp1240	Unsupervised Multiplex Graph learning with Complementary and Consistent Information
mmfp1246	Freq-HD: an Interpretable Frequency-based High-Dynamics Affective Clip Selection Method for in-the-Wild Facial Expression Recognition in Videos
mmfp1281	Clip Fusion with Bi-level Optimization for Human Mesh Reconstruction from Monocular Videos
mmfp1288	Rethinking Pseudo-Label-Based Unsupervised Person Re-ID with Hierarchical Prototype-based Graph
mmfp1292	Scalable Incomplete Multi-View Clustering with Structure Alignment
mmfp1293	Point-aware Interaction and CNN-induced Refinement Network for RGB-D salient object detection
mmfp1298	Federated Learning with Label-Masking Distillation
mmfp1301	Modality Profile - A New Critical Aspect to be Considered When Generating RGB-D Salient Object Detection Training Set
mmfp1308	Efficient Multi-View Graph Clustering with Local and Global Structure Preservation
mmfp1314	Frequency-based Zero-Shot Learning with Phase Augmentation
mmfp1327	Single-stage Multi-human Parsing via Point Sets and Center-based Offsets
mmfp1376	Multi-Spectral Image Stitching via Spatial Graph Reasoning
mmfp1382	Cooperative Colorization: Exploring Latent Cross-Domain Priors for NIR Image Spectrum Translation
mmfp1388	IRCasTRF: Inverse Rendering by Optimizing Cascaded Tensorial Radiance Fields, Lighting, and Materials From Multi-view Images
mmfp1423	Temporally Efficient Gabor Transformer for Unsupervised Video Object Segmentation
mmfp1461	ALEX: Towards Effective Graph Transfer Learning with Noisy Labels
mmfp1464	Federated Deep Multi-View Clustering with Global Self-Supervision
mmfp1528	StyleEDL: Style-Guided High-order Attention Network for Image Emotion Distribution Learning
mmfp1542	Multi-Frame Self-Supervised Depth Estimation with Multi-Scale Feature Fusion in Dynamic Scenes
mmfp1580	Unveiling the Power of CLIP in Unsupervised Visible-Infrared Person Re-Identification
mmfp1615	ScribbleVC: Scribble-supervised Medical Image Segmentation with Vision-Class Embedding
mmfp1643	CoMoSpeech: One-Step Speech and Singing Voice Synthesis via Consistency Model
mmfp1649	Lifelong Scene Text Recognizer via Expert Modules
mmfp1680	Graph based Spatial-temporal Fusion for Multi-modal Person Re-identification
mmfp1684	Noise-Robust Continual Test-Time Domain Adaptation
mmfp0070	Cultural Self-Adaptive Multimodal Gesture Generation Based on Multiple Culture Gesture Dataset
mmfp0084	PiPa: Pixel- and Patch-wise Self-supervised Learning for Domain Adaptative Semantic Segmentation
mmfp0089	DPNET: Dynamic Poly-attention Network for Trustworthy Multi-modal Classification
mmfp0169	DeNoising-MOT: Towards Multiple Object Tracking with Severe Occlussions

mmfp0175	DocDiff: Document Enhancement via Residual Diffusion Models
mmfp0266	Uncertainty-Guided Spatial Pruning Architecture for Efficient Frame Interpolation
mmfp0279	Scene Text Segmentation with Text-Focused Transformers
mmfp0344	Beyond First Impressions: Integrating Joint Multi-modal Cues for Comprehensive 3D Representation
mmfp0366	One-stage Low-resolution Text Recognition with High-resolution Knowledge Transfer
mmfp0369	PSNEA: Pseudo-Siamese Network for Entity Alignment between Multi-modal Knowledge Graphs
mmfp0391	Exploring Correlations in Degraded Spatial Identity Features for Blind Face Restoration
mmfp0399	Semantic-based Selection, Synthesis, and Supervision for Few-shot Learning
mmfp0418	Implicit Decouple Network for Efficient Pose Estimation
mmfp0465	Towards Robust Real-Time Scene Text Detection: From Semantic to Instance Representation Learning
mmfp0490	Diffusion-Augmented Depth Prediction with Sparse Annotations
mmfp0506	Unite-Divide-Unite: Joint Boosting Trunk and Structure for High-accuracy Dichotomous Image Segmentation
mmfp0533	Self-Contrastive Graph Diffusion Network
mmfp0580	All-in-one Multi-degradation Image Restoration Network via Hierarchical Degradation Representation
mmfp0653	NPF-200: A Multi-Modal Eye Fixation Dataset and Method for Non-Photorealistic Videos
mmfp0670	SkipStreaming: Pinpointing User-Perceived Redundancy in Correlated Web Video Streaming through the Lens of Scenes
mmfp0683	Efficient Parallel Multi-Scale Detail and Semantic Encoding Network for Lightweight Semantic Segmentation
mmfp0767	Patchmatch Stereo++: Patchmatch Binocular Stereo with Continuous Disparity Optimization
mmfp0787	UAVTracker: Fast multiple object tracking from UAV-captured videos based on optical flow
mmfp0904	Domain-irrelevant Feature Learning for Generalizable Pan-sharpening
mmfp0908	Dynamic Grouped Interaction Network for Low-Light Stereo Image Enhancement
mmfp0963	Learnable Graph Filter for Multi-view Clustering
mmfp0964	Pixel Adapter: A Graph-Based Post-Processing Approach for Scene Text Image Super-Resolution
mmfp1063	Object Part Parsing with Hierarchical Dual Transformer
mmfp1083	Learning High-frequency Feature Enhancement and Alignment for Pan-sharpening
mmfp1120	Multi-Modal and Multi-Scale Temporal Fusion Architecture Search for Audio-Visual Video Parsing
mmfp1130	DUSA: Decoupled Unsupervised Sim2Real Adaptation for Vehicle-to-Everything Collaborative Perception
mmfp1143	Pseudo Object Replay and Mining for Incremental Object Detection
mmfp1144	FedAA: Using Non-sensitive Modalities to Improve Federated Learning while Preserving Image Privacy
mmfp1233	Object Segmentation by Mining Cross-Modal Semantics
mmfp1254	AcFormer: An Aligned and Compact Transformer for Multimodal Sentiment Analysis
mmfp1313	DecenterNet: Bottom-Up Human Pose Estimation Via Decentralized Pose Representation
mmfp1368	Video Infringement Detection via Feature Disentanglement and Mutual Information Maximization
mmfp1408	Cross-modal Unsupervised Domain Adaptation for 3D Semantic Segmentation via Bidirectional Fusion-then-Distillation
mmfp1470	Where and How: Mitigating Confusion in Neural Radiance Fields from Sparse Inputs
mmfp1502	View while Moving: Efficient Video Recognition in Long-untrimmed Videos
mmfp1526	Where to Find Fascinating Inter-Graph Supervision: Imbalanced Graph Classification with Kernel Information Bottleneck
mmfp1558	Entropy Neural Estimation for Graph Contrastive Learning
mmfp1636	PBFormer: Capturing Complex Scene Text Shape with Polynomial Band Transformer
mmfp1716	CenterLPS: Segment Instances by Centers for LiDAR Panoptic Segmentation
mmfp1736	Incorporating Domain Knowledge Graph into Multimodal Movie Genre Classification with Self-Supervised Attention and Contrastive Learning
mmfp1834	Towards Flexible and Universal: A Novel Endpoint-based Framework for Vessel Structural Information Extraction
mmfp1898	Mask Again: Masked Knowledge Distillation for Masked Video Modeling
mmfp1903	Fine-grained Key-Value Memory Enhanced Predictor for Video Representation Learning
mmfp1929	PNT-Edge: Towards Robust Edge Detection with Noisy Labels by Learning Pixel-level Noise Transitions
mmfp1957	DTF-Net: Category-Level Pose Estimation and Shape Reconstruction via Deformable Template Field
mmfp1967	DANet: Multi-scale UAV Target Detection with Dynamic Feature Perception and Scale-aware Knowledge Distillation
mmfp2002	Cross-Modal Graph Attention Network for Entity Alignment
mmfp2006	Exploring Motion Cues for Video Test-Time Adaptation