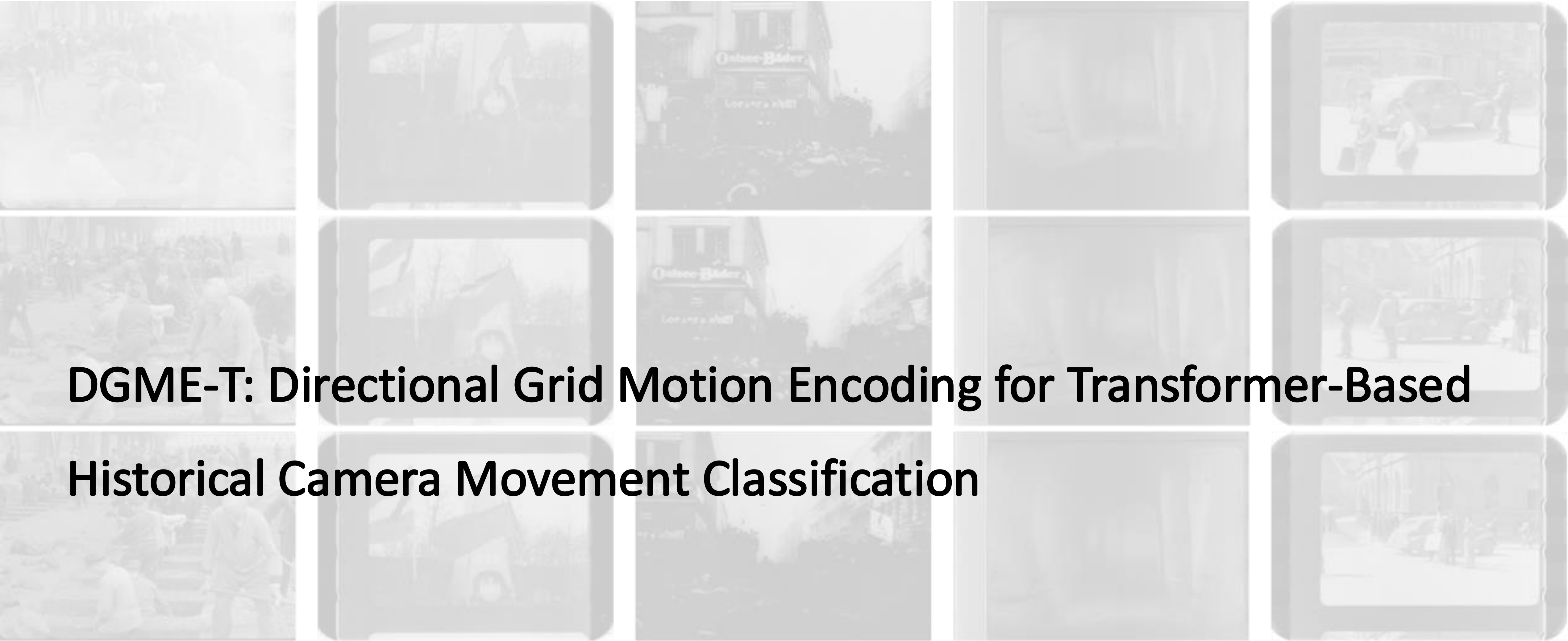


DGME-T: Directional Grid Motion Encoding for Transformer-Based Historical Camera Movement Classification

Tingyu Lin, Armin Dadras, Florian Kleber, and Robert Sablatnig

TU WIEN CVL Computer Vision Lab

FWF Österreichischer Wissenschaftsfonds

Supported by the Austrian Science Fund (FWF)
doc.funds.connect, under project grant no. DFH 37-N.

(a)

(b)

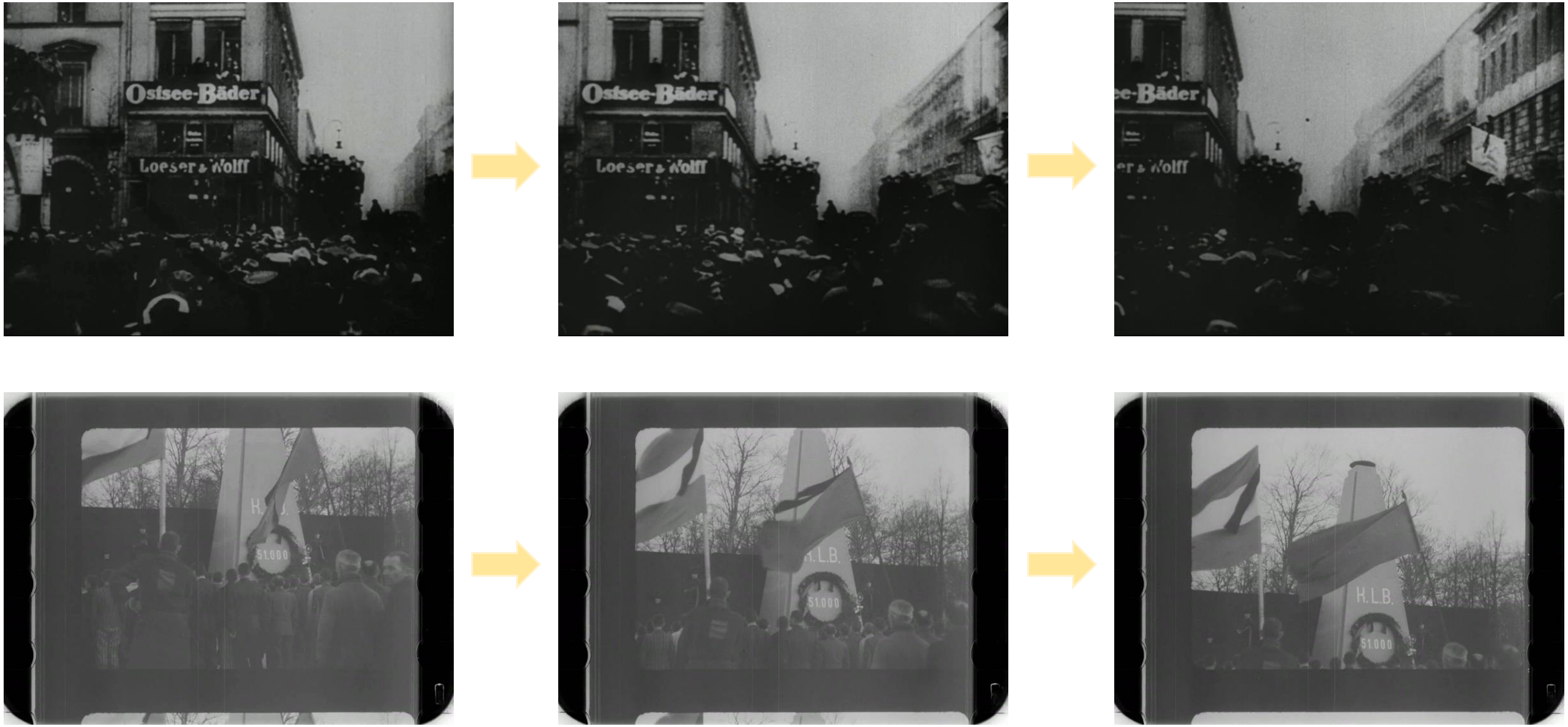


Fig. 1: Examples of basic camera motions, left-to-right panning (a) and bottom-to-top tilting (b).

Why Camera Movement Matters

- Camera movement is a core element of cinematic language, it shapes narrative rhythm and emotional tone. Understanding movements reveals directorial style and storytelling intent.
- For film scholars, automatic camera movement analysis enables quantitative studies of style. For archives and cultural heritage, it supports metadata enrichment, retrieval, and restoration workflows.

→ Reliable Camera Movement Classification (CMC) is a key step toward scalable video understanding.

Camera Movement Classification (CMC)

- Task: assign semantic motion labels (e.g., static, pan, tilt, zoom, track) to short video segments.
- Focuses on camera motion, not object motion, aims to capture how the camera itself moves.
- Applications: film style analysis, video retrieval, shot detection, and automatic editing tools.
- Recent advances use deep models (CNNs, Transformers) for robust feature learning from raw frames.

Modern Dataset

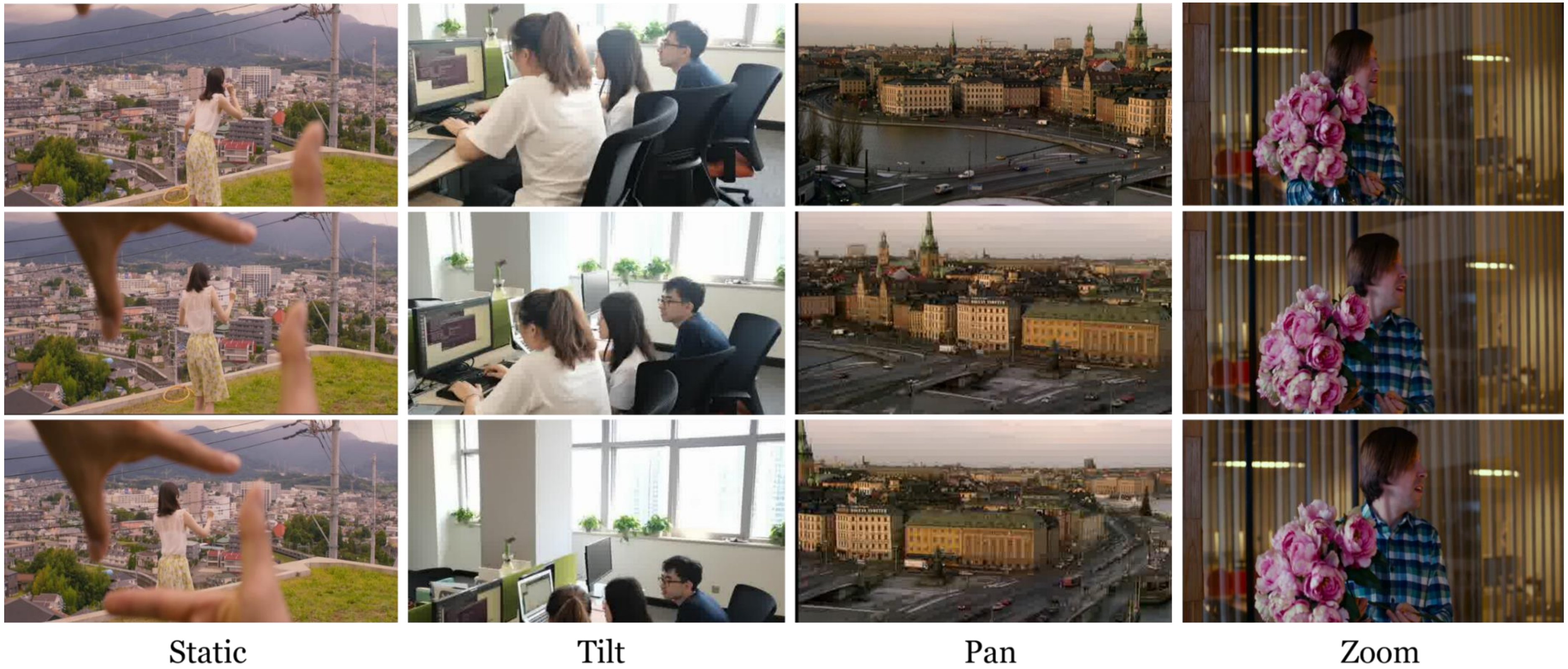


Fig. 2: Example frames from the modern training dataset showing clean, high–resolution video content.

Zeyu Chen, Yana Zhang, Lianyi Zhang, and Cheng Yang. 2021. Ro-textcnn based mul-move-net for camera motion classification. In 2021 IEEE/ACIS 20th International Fall Conference on Computer and Information Science (ICIS Fall). IEEE, 182–186.

Anyi Rao, Jiaze Wang, Linning Xu, Xuekun Jiang, Qingqiu Huang, Bolei Zhou, and Dahua Lin. 2020. A Unified Framework for Shot Type Classification Based on Subject Centric Lens. In The European Conference on Computer Vision (ECCV). Springer, 17–34.

Historical Footage

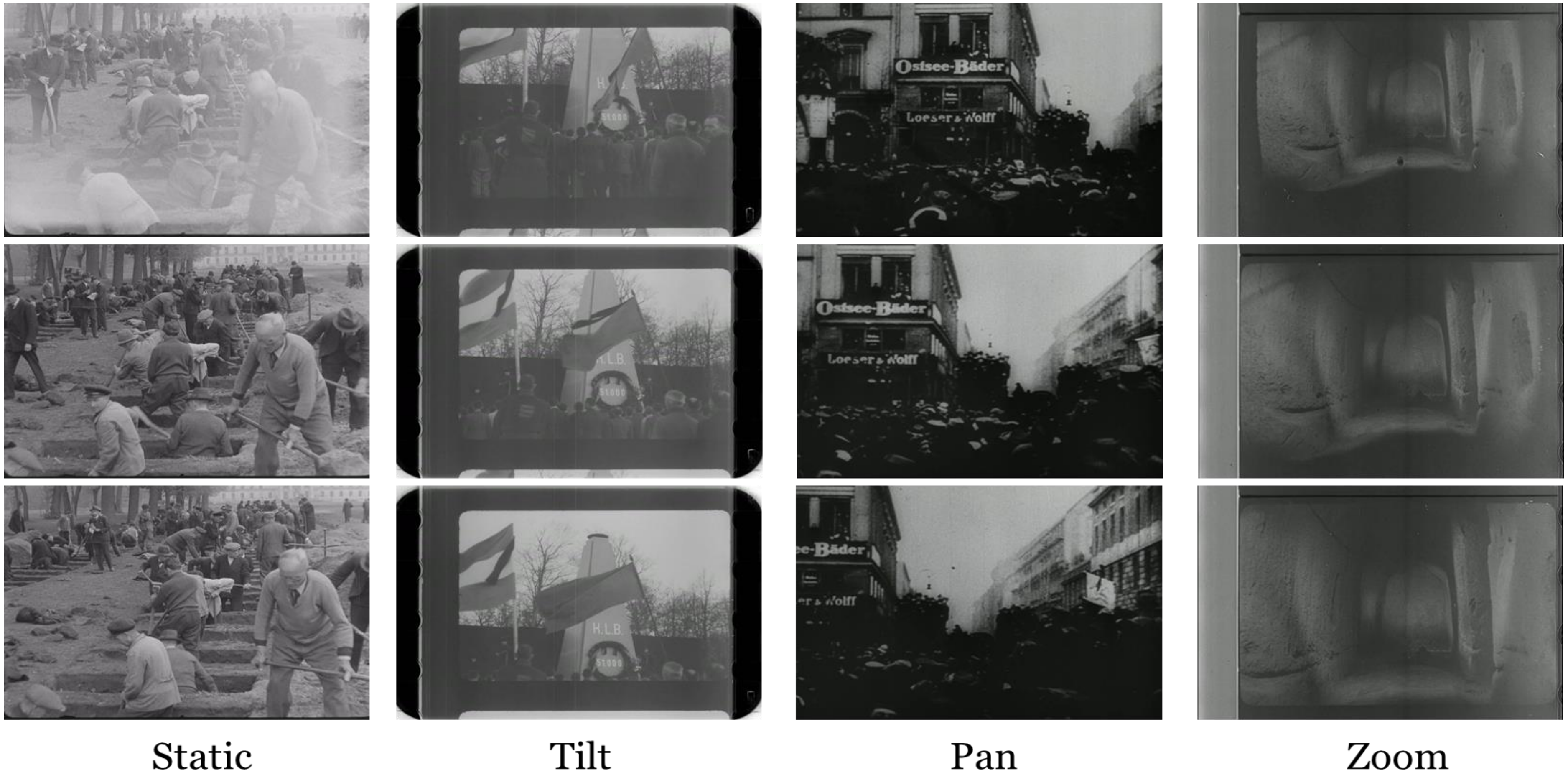


Fig. 3: Example frames from the HISTORIAN dataset illustrating typical visual degradation, blur, and low contrast encountered in archival footage.

Daniel Helm, Fabian Jogl, and Martin Kampel. "Historian: A Large-Scale Historical Film Dataset with Cinematographic Annotation". In: 2022 IEEE International Conference on Image Processing (ICIP). IEEE. 2022, pp. 2087–2091.

Historical Footage

Fig. 4: Example of a track camera movement from the HISTORIAN dataset. Frames are sampled every 20 frames to illustrate the motion.

Daniel Helm, Fabian Jogl, and Martin Kampel. “Historian: A Large-Scale Historical Film Dataset with Cinematographic Annotation”. In: 2022 IEEE International Conference on Image Processing (ICIP). IEEE. 2022, pp. 2087–2091.

The Challenge: Historical Footage

- Archival films often suffer from:
  - Noise, blur, missing frames, unstable exposure
  - Low contrast and digitization artifacts
- These degradations distort motion cues and break assumptions of modern CMC models trained on clean data.
- Historical datasets are small and imbalanced, so deep models tend to overfit and fail to generalize across domains.
- As a result, accuracy of modern CMC models drops sharply when applied to historical footage.

DGME-T

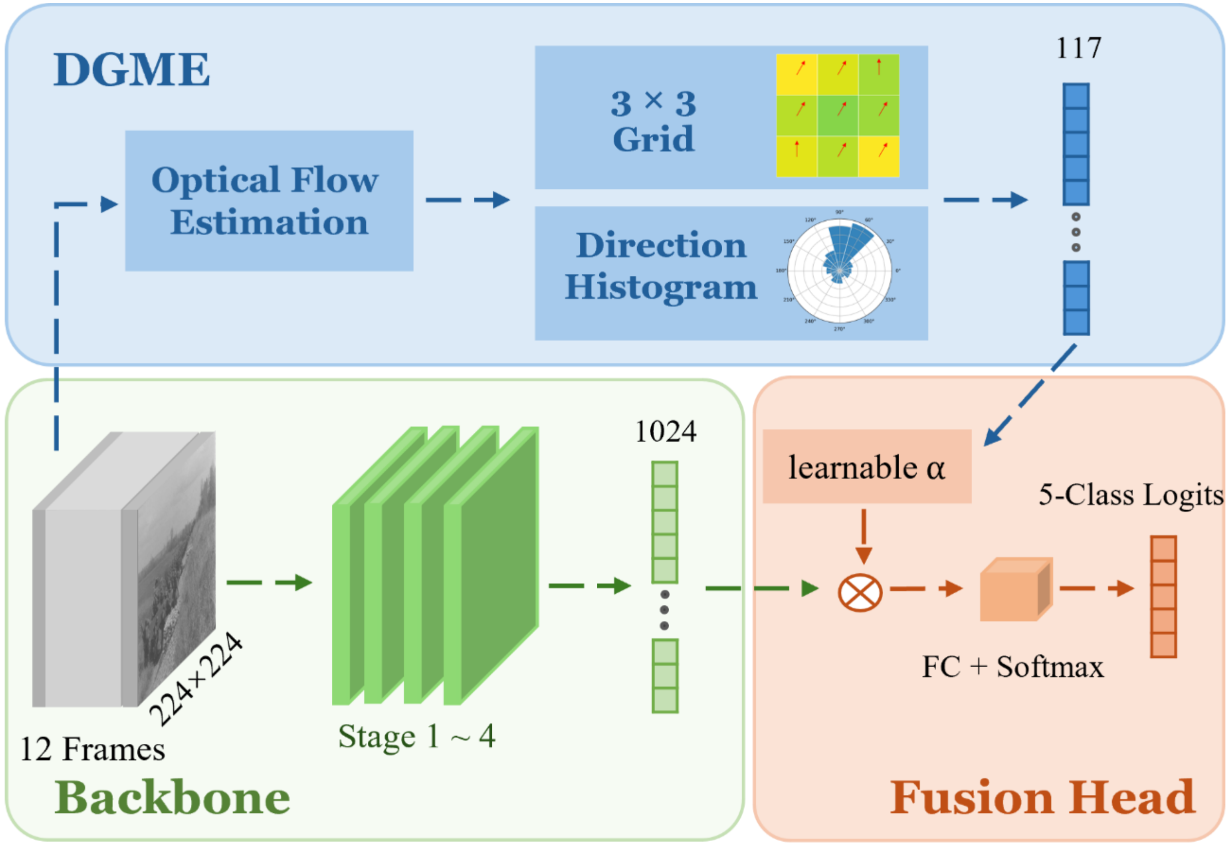


Fig. 5: Overall architecture of DGME-T, combining directional motion encoding with a Video Swin Transformer backbone.

- Key insight: Structured motion priors and transformer representations are complementary.
- Directional Grid Motion Encoding (DGME):captures local directional motion patterns through 3×3 grid histograms.
- Late Fusion:integrates DGME features with the Video Swin Transformer output via a learnable normalization layer.

Computer Vision Lab

Directional Grid Motion Encoding (DGME)

- Why DGME
  - Historical videos often lose clear motion cues. Deep features alone miss fine directional details. DGME brings explicit, interpretable motion evidence into the model.
- How It Works
  - Compute dense optical flow between frames. Divide frame into a 3×3 grid. Build weighted directional histograms per cell (+ static bin). Concatenate and normalize → compact motion descriptor.
- Why It Matters
  - Captures localized motion patterns that deep networks overlook. Improves recognition of subtle or degraded movements. Adds minimal cost, yet enhances directional robustness.

Hasan, Muhammad Abul, Min Xu, Xiangjian He, and Changsheng Xu."CAMHID: Camera Motion Histogram Descriptor and Its Application to Cinematographic Shot Classification."In: IEEE Transactions on Circuits and Systems for Video Technology 24, no. 10 (2014): 1682–1695.

Global 12-direction histograms across classes

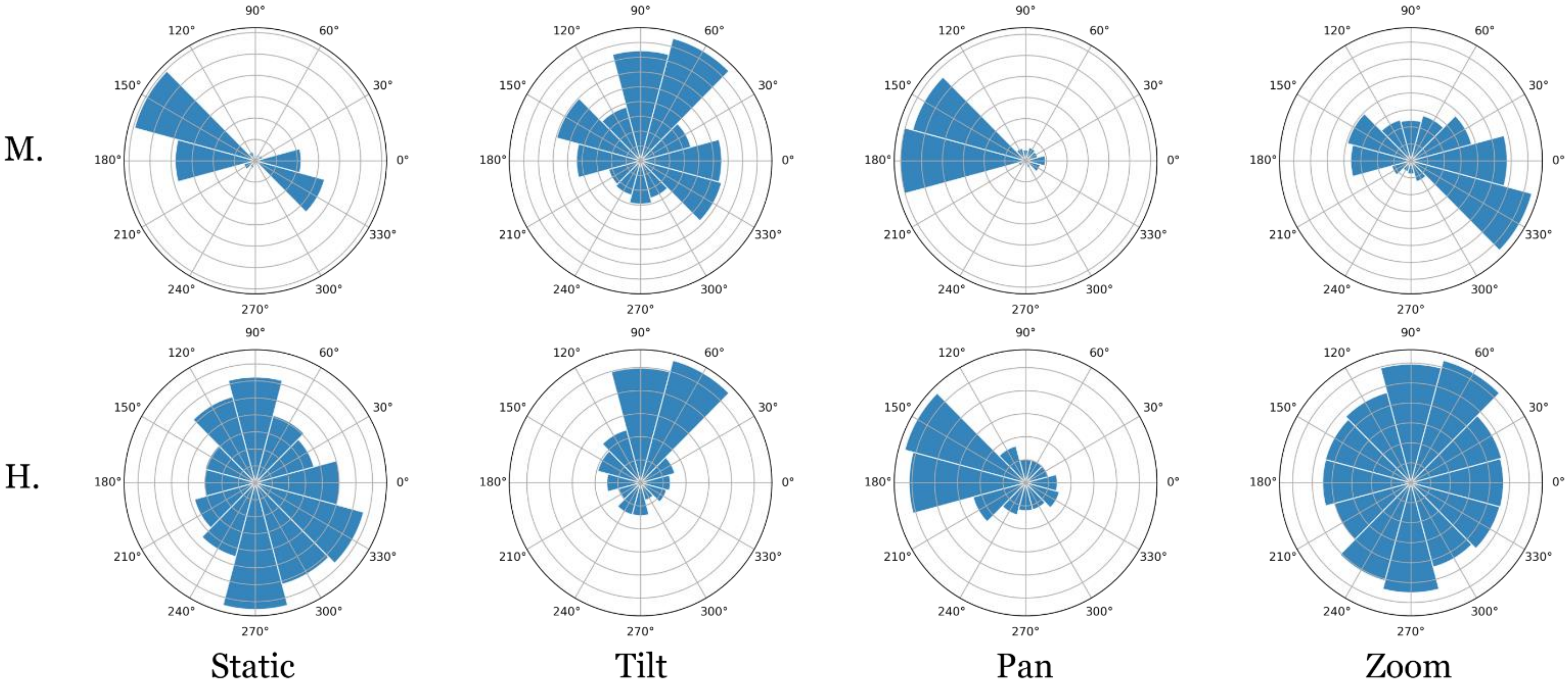


Fig. 5: Global 12-direction rose diagrams for four movement classes.

DGME 3×3 grid visualisation

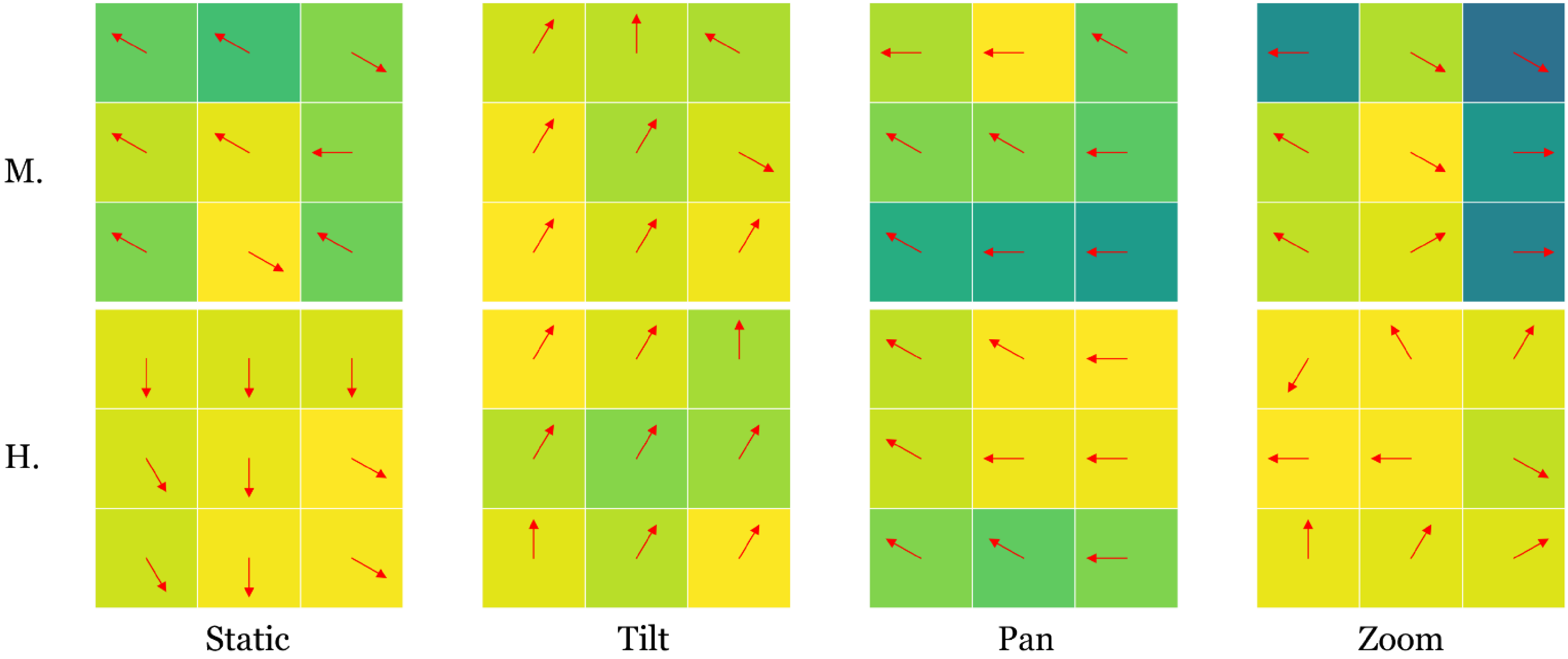


Fig. 6: DGME 3X3 grid visualisation. Cell colour encodes motion magnitude, arrows indicate the dominant direction in each cell.

Datasets

Multiple datasets exist for CMC, but vary widely in source, scale, and label definitions, limiting cross-domain transferability.

Dataset	Video Source	Scale (Shots / Videos)	Types
HISTORIAN	WWII archival films	838 movements / 98 films	8
MovieShots	Modern movie trailers	46,857 shots / 7,858 videos	4
MOVE-SET	Multi-domain video content	100K+ frame pairs / 448 videos	9

Datasets

Multiple datasets exist for CMC, but vary widely in source, scale, and label definitions, limiting cross-domain transferability.

Dataset	Camera Movement Types
HISTORIAN	pan, tilt, track, truck, dolly, zoom, pedestal, pan_tilt
MovieShots	static, motion, push, pull
MOVE-SET	static, up, down, left, right, zoom in, zoom out, rotate left, rotate right

Datasets and Label Alignment

- Modern domain: combined from MOVE-SET and MovieShots.
- Historical domain: HISTORIAN, containing annotated WWII footage.
- Labels were unified into four modern and five historical classes:static, tilt, pan, zoom (+ track only in HISTORIAN).
- This alignment enables cross-domain evaluation and consistent training.

Table 1: Revised HISTORIAN dataset sample distribution.

Class	Static	Tilt	Pan	Zoom	Track
Source	new	tilt+pedestal	pan+truck	zoom+dolly	track
Count	82	116	304	77	252

Computer Vision Lab

TU WIEN

Cross-Domain Transfer and Fine-Tuning Strategy

Table 2: Per-class precision (P), recall (R) and F_1 on HISTORIAN. All numbers are percentages.

Class	Kinetics-only			Modern-Historical		
	P	R	F_1	P	R	F_1
Static	88.24	88.24	88.24	93.75	88.24	90.91
Tilt	81.82	75.00	78.26	94.74	75.00	83.72
Pan	75.68	91.80	82.96	84.06	95.08	89.23
Zoom	85.71	37.50	52.17	81.82	56.25	66.67
Track	75.51	72.55	74.00	75.93	80.39	78.10
Macro avg.	81.39	73.02	75.13	86.06	78.99	81.72

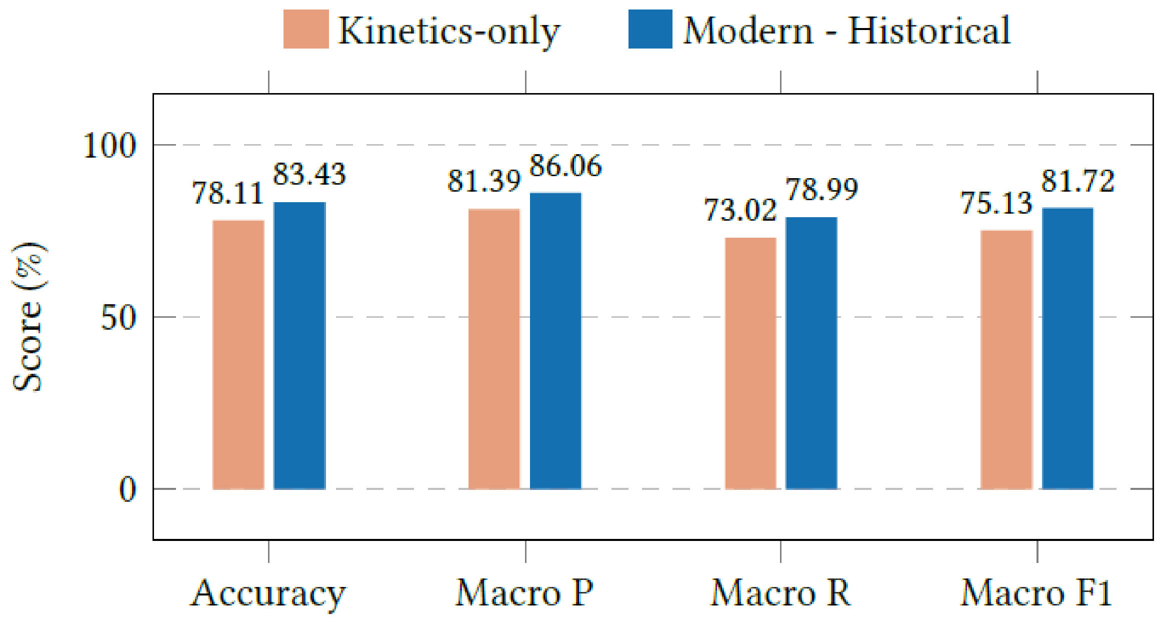


Figure 7: Macro-level performance comparison for cross-domain transfer.

- Modern pre-training improves HISTORIAN performance by +5% accuracy and +6% macro F1.
- Especially strong gains for tilt (+13% precision) and zoom (+14% F1) classes.

Computer Vision Lab

Model Comparison Across Domains

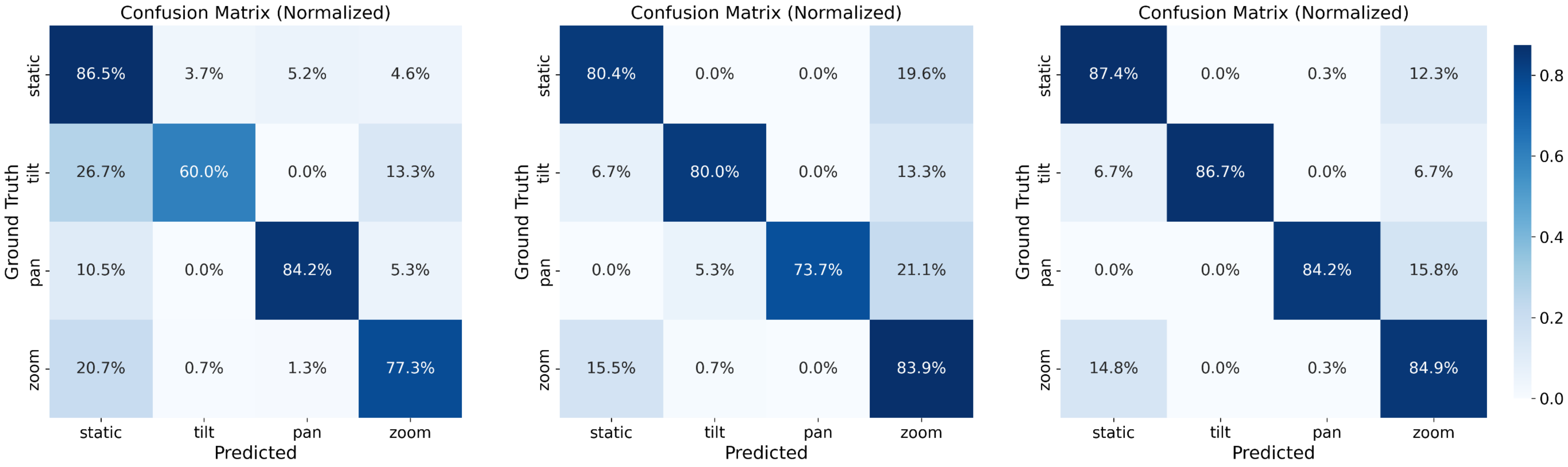
Table 3: Overall performance of three models on modern and historical datasets.

Model	Modern Dataset Acc (%)	Modern Dataset F_1 (%)	HISTORIAN Dataset Acc (%)	HISTORIAN Dataset F_1 (%)
CAMHID (DGME-only)	81.63	68.05	55.62	54.22
Video Swin	81.78	82.08	83.43	81.72
DGME-T (Ours)	**86.14**	**87.81**	**84.62**	**82.63**

- On modern data:
  - +5.7 F1 improvement over Video Swin → better pan/tilt discrimination.
- On historical data:
  - DGME-T still achieves highest accuracy (84.6%) and macro-F1 (82.6%).

- Handcrafted motion (CAMHID) fails under noise;DGME-T combines semantic + directional cues effectively.

Model Comparison Across Domains



(a) CAMHID – Modern

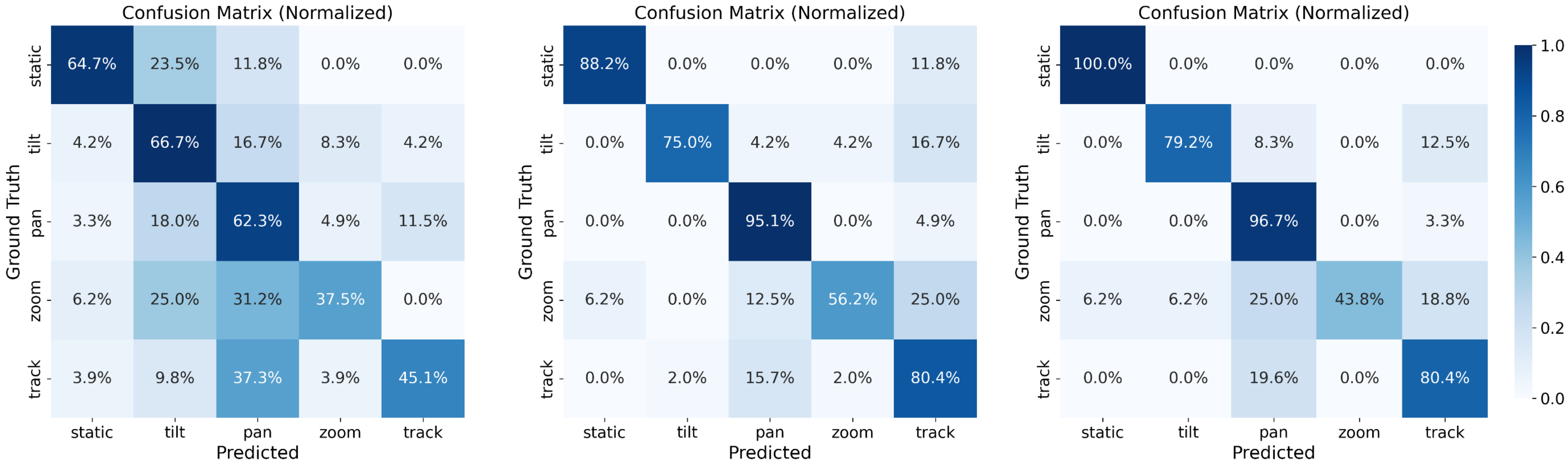
(b) Video Swin – Modern

(c) DGME-T – Modern

Figure 7: Confusion matrices for three models on modern datasets

Computer Vision Lab

Model Comparison Across Domains



(d) CAMHID – HISTORIAN (e) Video Swin – HISTORIAN (f) DGME-T – HISTORIAN

Figure 7: Confusion matrices for three models on HISTORIAN

Conclusion

- Introduced DGME-T, combining Directional Grid Motion Encoding with a Video Swin Transformer.
- Achieves consistent gains on both modern and historical footage.
- Macro-F1: 82.1 → 87.8 (modern), 81.7 → 82.6 (historical).
- Demonstrates that structured motion priors and attention-based features are complementary.

Next Steps

- Explore alternative optical flow estimators and motion descriptors.
- Investigate mid-level or temporal fusion strategies.
- Extend cross-domain studies to other film periods or restoration contexts.
- Apply DGME-T to heritage digitization pipelines for automatic metadata enrichment.

Computer Vision Lab

Thank you!

Tingyu Lin

Survey

CMC methods evolved from rule-based heuristics to deep spatiotemporal models. We summarize key approaches across input types and model designs.

Table 4: Comparison of representative CMC methods.

Method	Model Type	Input Features	Types
Wang & Cheong	Rule-based + MRF	Optical flow, motion entropy, attention maps	7
CAMHID	Rule-based + SVM	Macroblock motion vectors	4
2D Histogram	Rule-based + matching	2D histograms of flow direction and magnitude	10
SGNet	Multi-branch CNN	RGB, saliency, segmentation	4
MUL-MOVE-Net	CNN + BiLSTM	Optical flow histograms	9
Petrogianni et al.	CNN + LSTM / SVM	Low-level visual statistics	10
LWSRNet	Lightweight 3D CNN	RGB, flow, saliency, segmentation	8

Method

Table 5: Performance of each model on the HISTORIAN validation set.

Model	Top-1 Accuracy (%)	Top-2 Accuracy (%)	Weighted F1 (%)
C3D	64.20	81.48	59.16
R(2+1)D	48.15	64.20	37.28
TSN	50.62	75.31	40.19
I3D	74.07	77.78	69.50
Video Swin	**80.25**	**87.65**	**76.24**