



Peer Community Journal

Section: Archaeology

Research article

Published
2026-05-27

Cite as

David Roqui, Adèle Cormier, Nistor Grozavu and Ann Bourges (2026) *A multimodal approach to heritage preservation in the context of climate change*, Peer Community Journal, 6: e48.

Correspondence
roquidavid@yahoo.fr

Peer-review
Peer reviewed and recommended by PCI Archaeology,
<https://doi.org/10.24072/pci.archaeo.100708>



This article is licensed under the Creative Commons Attribution 4.0 License.

A multimodal approach to heritage preservation in the context of climate change

David Roqui^{1,2,3}, Adèle Cormier^{3,4,2}, Nistor Grozavu¹, and Ann Bourges³

Volume 6 (2026), article e48

<https://doi.org/10.24072/pcjournal.723>

Abstract

Cultural heritage sites face accelerating degradations due to climate change, yet traditional monitoring relies on unimodal analysis (visual inspection or environmental sensors alone) that fails to capture the complex interplay between environmental stressors and material deterioration. We propose a lightweight multimodal architecture that fuses sensor data (temperature, humidity) with visual imagery to predict degradation severity at heritage sites. Our approach adapts PerceiverIO with two key innovations: (1) simplified encoders (64D latent space) that prevent overfitting on small datasets (37 samples for training, 555 with data augmentation; 13 for validation, and 13 for testing), and (2) Adaptive Barlow Twins loss that encourages modality complementarity rather than redundancy. On data from Strasbourg Cathedral, our model achieves 76.9% accuracy and 77.0% weighted-F1 score on the test set, a 43% improvement over standard multimodal architectures (VisualBERT, Transformer) and 25% over vanilla PerceiverIO. Ablation studies reveal that sensor-only achieves 61.5% while image-only reaches 46.2%, confirming successful multimodal synergy. A systematic hyperparameter study identifies an optimal moderate correlation target ($\tau = 0.3$) that balances alignment and complementarity, achieving 69.2% accuracy compared to other τ values ($\tau = 0.1/0.5/0.7$: 53.8%, $\tau = 0.9$: 61.5%). This work demonstrates that architectural simplicity combined with contrastive regularization enables effective multimodal learning in data-scarce heritage monitoring contexts, providing a foundation for AI-driven conservation decision support systems.

¹Laboratoire ETIS, UMR 8051, CY Cergy Paris Université - ENSEA - CNRS, Cergy-Pontoise, France, ²Fondation des sciences du patrimoine (FSP), Paris, France, ³Centre de recherche et de restauration des musées de France (C2RMF), UMR 8247 IRCP-PCMTH, Paris, France, ⁴Epitopos, Paris, France

Peer Community Journal is a member of the
Centre Mersenne for Open Scientific Publishing
<http://www.centre-mersenne.org/>

e-ISSN 2804-3871



Introduction

Climate change accelerates the degradations of cultural heritage worldwide through temperature fluctuations, humidity cycles, and extreme weather events. Traditional conservation relies on periodic expert inspections, a reactive approach insufficient for the pace of climate-driven deterioration. Automated monitoring systems could enable proactive interventions, yet heritage preservation presents unique machine learning challenges. Datasets rarely exceed one hundred samples due to limited site access, expensive expert annotation, and slow degradation timescales spanning years.

While climate change provides the broader long-term motivation for automated heritage monitoring, this work presents a proof-of-concept on data from a single campaign (T0). The approach establishes a foundation for future temporal analysis as multi-year monitoring data (T1, T2, etc.) becomes available.

Degradation results from complex interactions between environmental stressors and material properties that visual inspection alone cannot capture. Temperature variations cause thermal expansion weakening structural bonds, while humidity drives salt crystallization within porous materials. This motivates multimodal approaches fusing sensor data with weathering monitoring through imagery. However, existing architectures like VisualBERT (Li et al., 2019), UNITER (Chen et al., 2020), and FLAVA (Singh et al., 2022) achieve strong performance on large-scale vision-language benchmarks but fail on specialized small-scale tasks. Pre-trained representations from general images do not transfer to scientific heritage imaging, while high parameter counts cause severe overfitting on limited training sets.

We propose a lightweight multimodal architecture adapted for data-scarce heritage monitoring. Our approach modifies PerceiverIO (Jaegle et al., 2022) through two key innovations. First, we replace complex encoders with simple linear projections, reducing parameters to match dataset size and prevent memorization. Second, we introduce Adaptive Barlow Twins loss that encourages modality complementarity rather than redundancy. Unlike standard fusion methods promoting identical representations, our partial correlation target preserves modality-specific information while maintaining semantic coherence.

We validate this approach on Strasbourg Cathedral monitoring data combining environmental sensors with surface imagery across five degradation classes. Through systematic ablation studies and hyperparameter analysis, we investigate how architectural simplification affects generalization, how modalities contribute individually versus combined, and what balance between alignment and complementarity optimizes performance.

The paper proceeds as follows. Section 2 reviews related work. Section 3 details our architecture and loss formulation. Section 4 describes the dataset and evaluation protocol. Sections 5 and 6 present results and discussion. Section 7 concludes with future directions.

Related Work

Multimodal Learning Architectures

Early multimodal approaches relied on modality-specific feature extractors followed by concatenation (Ngiam et al., 2011a). Convolutional neural networks for images (Bengio and LeCun, 1995) and recurrent networks for sequences (Hochreiter and Schmidhuber, 1997) processed each modality independently before late fusion through fully connected layers. However, this

strategy fails to capture cross-modal interactions during feature learning, limiting representational power.

The transformer architecture (Vaswani et al., 2018) revolutionized multimodal learning by enabling attention-based fusion. VisualBERT (Li et al., 2019) extended BERT (Turc et al., 2019) to vision-language tasks through co-attentional layers, achieving strong performance on visual question answering and image captioning. UNITER (Chen et al., 2020) and FLAVA (Singh et al., 2022) further improved cross-modal alignment through contrastive pre-training on large-scale image-text pairs. Vision Transformers (Dosovitskiy et al., 2021) demonstrated that pure attention mechanisms could match or exceed convolutional architectures on image classification when sufficient training data is available.

Despite their success on large-scale benchmarks, these models face critical limitations in specialized domains. First, pre-training on general vision-language corpora does not transfer effectively to scientific imaging modalities like multispectral sensors or microscopy. Second, model complexity requires datasets with tens of thousands of samples to avoid overfitting, far exceeding typical heritage monitoring budgets. Third, these architectures assume semantic alignment between modalities, whereas our task requires preserving complementarity between environmental sensors and visual evidence.

Fusion Strategies for Heterogeneous Modalities

The choice of fusion strategy critically impacts multimodal performance. Early fusion concatenates raw inputs before processing (Ngiam et al., 2011b), enabling joint feature learning but increasing dimensionality and computational cost. Late fusion combines predictions from modality-specific models (Karpathy et al., 2014), preserving specialization but missing cross-modal interactions during training. Intermediate fusion balances these trade-offs through hierarchical integration at multiple network depths (Poria et al., 2017).

Perceiver (Jaegle et al., 2021) introduced a paradigm shift by mapping diverse input modalities to a shared latent space through iterative cross-attention. This approach handles variable-sized inputs and scales linearly with input length rather than quadratically like standard transformers. PerceiverIO (Jaegle et al., 2022) extended this framework with flexible output decoders, enabling task-specific predictions while maintaining architectural generality. However, the original Perceiver design targets large-scale pre-training scenarios and requires adaptation for small-data regimes.

Self-Supervised Learning for Multimodal Representations

Recent work explores contrastive objectives for learning aligned multimodal representations without explicit labels. CLIP (Radford et al., 2021) trains vision and language encoders to maximize similarity between corresponding image-text pairs while minimizing similarity for mismatched pairs. Barlow Twins (Zbontar et al., 2021) reduces redundancy between augmented views by decorrelating their representations, avoiding collapse without requiring negative samples.

These methods assume modalities provide redundant views of the same semantic content. Heritage monitoring violates this assumption: sensors capture environmental causes while images reveal material effects. Our work adapts Barlow Twins from view-invariance to modality-complementarity, encouraging decorrelation rather than alignment.

Machine Learning for Heritage Preservation

AI applications in cultural heritage have primarily focused on digital reconstruction and damage detection. Generative adversarial networks restore degraded artworks (Elgammal et al., 2017), while convolutional networks detect cracks in historical structures from visual inspection (Dais et al., 2021). However, these approaches operate on single modalities and ignore environmental context.

Recent work explores multimodal heritage monitoring by combining visual surveys with climate data. Cabral et al. (2020) fuse thermal imaging with structural sensors for building assessment but rely on large annotated datasets (1000+ samples) unavailable for most sites. Grilli and Remondino (2019) integrate photogrammetry with environmental logging yet analyze modalities separately rather than jointly. While these works combine sensors with imagery, they require large datasets and analyze modalities separately. To our knowledge, ours is the first work to address joint multimodal fusion under extreme data scarcity ($n < 100$) through complementarity-driven learning for heritage monitoring.

Positioning of Our Work

This work is at the intersection of three research areas: small-data deep learning, contrastive multimodal fusion, and heritage science. We adapt PerceiverIO for data-scarce scenarios through architectural simplification, drawing inspiration from network pruning (Han et al., 2015) and knowledge distillation (Hinton et al., 2015) that demonstrate smaller models can match or exceed larger ones when training data is limited. Our Adaptive Barlow Twins loss extends contrastive learning from view-invariance to modality-complementarity, addressing the gap between existing self-supervised methods and heterogeneous sensor-image fusion. Finally, we provide a benchmark of state-of-the-art multimodal architectures on heritage monitoring data, establishing baselines for future research in this domain.

Dataset

Collected data comes from three French heritage sites (Bibracte archaeological site, Strasbourg Cathedral, and the Saint-Pierre Chapel) as part of a comprehensive heritage monitoring program (Cormier et al., 2025). These data are currently divided into two modalities : continuous text data from sensors and images punctuously collected on sites. Several data collection campaigns have already been conducted and occur every six months to monitor weathering evolution. Images are transformed into weathering maps with different layers. It is important to note that for this first article, data used are the first campaign (T0) data and the T0 data from Strasbourg Cathedral. However, other campaigns will take place, allowing for the creation of T1, T2, etc., improving the model's efficiency. Collected data from climatic sensors have good quality, with data collected regularly and without missing data or outlier values. Figures 2 to 4 show examples of what the sensor and image modality data may look like.

Climatic continuous data

This dataset from climatic and crack sensors represents the text modality. On the three sites, 16 thermohygrometer-sensors continuously record three parameters every hour : temperature (°C), relative humidity (%) and surface temperature (°C). The data is sent directly to a platform connected to the IoT as represented in Figure 1. Other sensors are related to the analysis of the

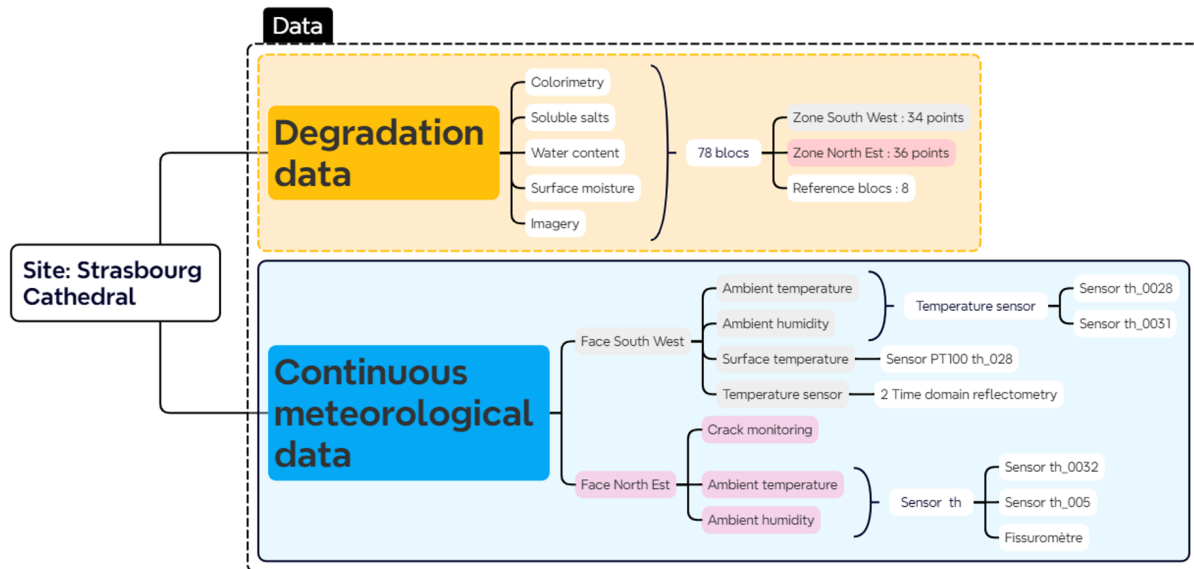


Figure 1 - Diagram of the dataset

monument's condition, such as crack meters or moisture content sensors. All data is collected in tables and then processed by an automated processing tool. Using this tool, a matrix of climatic metrics (statistics, number of dewpoint cycles, number of freeze-thaw cycles...) is extracted between two dates, as shown in 1. These matrices will be implemented in the model along with other image data from on-site alteration diagnostics.

The 28 sensor metrics include: mean, standard deviation, minimum, and maximum temperature (4 features); mean, standard deviation, minimum, and maximum relative humidity (4 features); mean and standard deviation of surface temperature (2 features); number of freeze-thaw cycles; number of dewpoint cycles; number of days above 90% relative humidity; cumulative thermal stress ($^{\circ}\text{C}\cdot\text{hours}$); humidity variance; diurnal temperature range; and 14 derivative metrics capturing temporal patterns. All metrics are computed over a 6-month aggregation window preceding each inspection campaign (October 2023 to April 2024 for T0). Features are normalized using z-score standardization: $z = (x - \mu) / \sigma$ where μ and σ are computed on the training set and applied consistently to validation and test splits.

Periodic imaging data

This image modality is composed of the different images taken during our data collection campaigns. These are scientific images captured using various acquisition modes, which are: direct light, grazing light, semi-grazing light, ultraviolet, and infrared, and thermogram taken with a thermal imaging camera. On site analysis such as colorimetry or complement these imaging campaigns. This leads to the creation of alteration maps, made on a drawing software. Each layer corresponds to an alteration pattern as defined in the ICOMOS glossary (Vergès-Belmin et al., 2011), as presented on 2. Detailed weathering characterization protocols and data collection methodology are described in (Cormier et al., 2025). Images are processed through a pre-trained Vision Transformer (ViT-B/16) encoder to extract 512-dimensional embeddings from the final [CLS] token. When multiple images exist for a single block (different angles), embeddings are averaged to produce a single 512D representation. A visual transformer is used to extract information from these images. The transformer architecture is particularly suited for processing

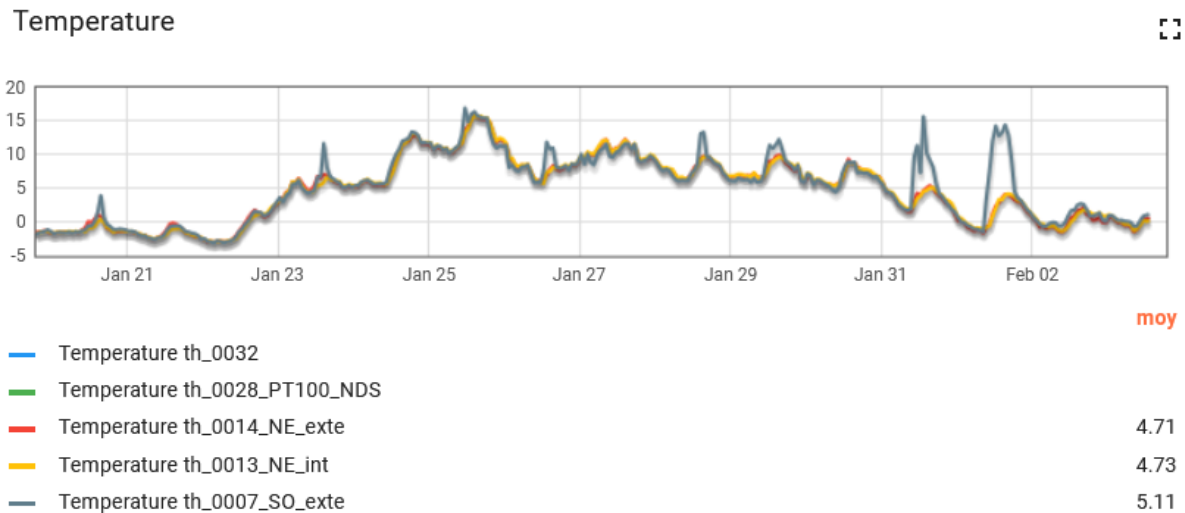


Figure 2 – Example of sensor data between two dates

this type of data because the images used can be complex in terms of information and links. Additionally, an image fusion is applied via the average of values for cases where a block, for example, is captured from different angles.

Figure 3 shows representative surface conditions.



Figure 3 – Image modality example

Final dataset

The final dataset used for this first round of experimentation is made only with Notre-Dame of Strasbourg site. It brings together the image and sensor modalities only for data collected at T0 in April 2024. It is a dataset composed of 70 data rows, with 4 rows where images are missing. For the 4 samples with missing images, the image encoder receives zero-padded inputs of the same dimensionality (512D), allowing the model to rely solely on sensor information through the cross-attention mechanism. After excluding samples with incomplete sensor data, 63 samples remain for experiments (37 train, 13 validation, 13 test). Additionally, there are 8 data rows that

come from control blocks, which will allow us to have a comparison point with the various blocks of the monuments.

The five degradation classes represent increasing severity: Class 0 (no visible degradation), Class 1 (minor discoloration), Class 2 (moderate surface alteration), Class 3 (significant material loss or deep cracks), Class 4 (severe structural degradation requiring urgent intervention). Classification follows the ICOMOS glossary criteria adapted for the site context.

Dataset Statistics

Table 1 – Dataset composition. Limited sample size reflects the challenge of expert-annotated heritage monitoring.

Split	N	Sensor dim	Image dim	Classes
Train	37	28	512	5
Validation	13	28	512	5
Test	13	28	512	5
Total	63	28	512	5

Train/val/test split follows 37/13/13 ratio with stratification by degradation class to ensure balanced representation. The limited test size ($n=13$) necessitates our 10-seed ensemble protocol for robust evaluation.

Methodology

We propose a lightweight multimodal architecture for heritage degradation assessment that adapts PerceiverIO (Jaegle et al., 2022) for small-scale datasets through two key features: (1) simplified encoders with regularization, and (2) Adaptive Barlow Twins loss that encourages modality complementarity rather than redundancy. Figure 4 illustrates the overall architecture.

Sensor and image data are encoded separately through lightweight linear projections (64D latent space), then fused via cross-attention. The Adaptive Barlow Twins loss encourages complementary representations, while the classification head predicts degradation severity.

Architectural Design

Modality-Specific Encoders. Given sensor data $\mathbf{s} \in \mathbb{R}^{d_s}$ and image features $\mathbf{i} \in \mathbb{R}^{d_i}$, we project each modality into a shared latent space of dimension $d_{\text{latent}} = 64$:

$$(1) \quad \mathbf{z}_s = \text{LN}(\text{Dropout}(\text{ReLU}(W_s \mathbf{s} + \mathbf{b}_s))),$$

$$(2) \quad \mathbf{z}_i = \text{LN}(\text{Dropout}(\text{ReLU}(W_i \mathbf{i} + \mathbf{b}_i))),$$

where $W_s \in \mathbb{R}^{64 \times d_s}$ and $W_i \in \mathbb{R}^{64 \times d_i}$ are learnable projections and dropout rate $p = 0.4$. where $W_s \in \mathbb{R}^{d_{\text{latent}} \times d_s}$ and $W_i \in \mathbb{R}^{d_{\text{latent}} \times d_i}$ are learnable projection matrices. We apply dropout ($p = 0.4$) to prevent overfitting on our limited training set ($n=70$ raw, 63 used after cleaning, 37 for training).

Unlike PerceiverIO Classic which uses full Perceiver encoders (128D latents, 3 self-attention blocks), our simplified projections reduce parameters from 50M to 12M while improving generalization. This aligns with sample complexity theory: model capacity should scale with dataset size (Shalev-Shwartz and Ben-David, 2014).

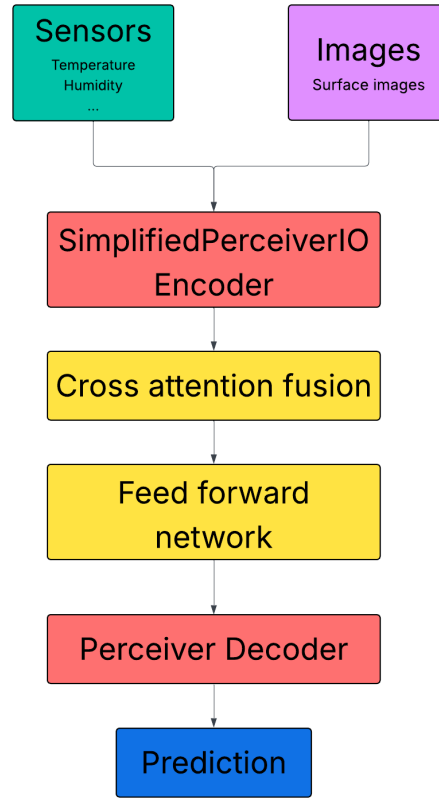


Figure 4 – Architecture overview

Cross-Attention Fusion. We fuse modality-specific representations using multi-head cross-attention:

$$(3) \quad \mathbf{z}_{\text{fused}} = \text{CrossAttn}(\mathbf{z}_s, \mathbf{z}_i, \mathbf{z}_i) + \mathbf{z}_s,$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_{\text{latent}}}}\right) V,$$

$$(4) \quad Q = W_Q \mathbf{z}_s, \quad K = W_K \mathbf{z}_i, \quad V = W_V \mathbf{z}_i,$$

$$(5) \quad \mathbf{z}_{\text{out}} = \text{FFN}(\mathbf{z}_{\text{fused}}) + \mathbf{z}_{\text{fused}},$$

$$(6) \quad \text{FFN}(\mathbf{x}) = W_2 \text{ReLU}(W_1 \mathbf{x} + b_1) + b_2.$$

where

$$\begin{aligned} \mathbf{z}_s, \mathbf{z}_i &\in \mathbb{R}^{d_{\text{latent}}} \text{ (sensor/image embeddings; typically } d_{\text{latent}} = 64), \\ W_Q, W_K, W_V &\in \mathbb{R}^{d_{\text{latent}} \times d_{\text{latent}}}, \quad W_1 \in \mathbb{R}^{d_{\text{ff}} \times d_{\text{latent}}}, \quad W_2 \in \mathbb{R}^{d_{\text{latent}} \times d_{\text{ff}}}, \\ b_1 &\in \mathbb{R}^{d_{\text{ff}}}, \quad b_2 \in \mathbb{R}^{d_{\text{latent}}}. \end{aligned}$$

Classification Head. The final prediction over $K = 5$ classes is

$$(7) \quad \hat{\mathbf{y}} = \text{softmax}(W_{\text{out}} \text{ReLU}(W_{\text{hid}} \mathbf{z}_{\text{out}})).$$

where $\hat{\mathbf{y}} \in \mathbb{R}^K$ represents predicted probabilities over $K = 5$ degradation classes.

Adaptive Barlow Twins Loss

Motivation. Standard multimodal fusion approaches (concatenation, element-wise operations) implicitly assume modalities provide redundant information. For heritage monitoring, this is sub-optimal: sensors capture environmental stressors (temperature, humidity) while images reveal visual manifestations (discoloration, cracks). We hypothesize that explicitly encouraging modality complementarity will improve generalization.

We adapt Barlow Twins (Zbontar et al., 2021), originally designed for self-supervised learning with augmented views, to multimodal fusion. The key modification is to replace the identity target (full correlation) with a partial correlation target that preserves modality-specific information.

Mathematical Formulation. Given a batch of sensor latents $\{\mathbf{z}_s^{(i)}\}_{i=1}^N$ and image latents $\{\mathbf{z}_i^{(i)}\}_{i=1}^N$, we compute the cross-correlation matrix:

$$(8) \quad C = \frac{1}{N} \sum_{i=1}^N \mathbf{z}_s^{(i)} (\mathbf{z}_i^{(i)})^T$$

where $\mathbf{z}_s^{(i)}$ and $\mathbf{z}_i^{(i)}$ are standardized representations:

$$(9) \quad \mathbf{z}_s^{(i)} = \frac{\mathbf{z}_s^{(i)} - \mathbb{E}[\mathbf{z}_s]}{\sqrt{\text{Var}[\mathbf{z}_s] + \epsilon}}$$

The Barlow Twins loss consists of two terms:

1. Diagonal term (partial alignment):

$$(10) \quad \mathcal{L}_{\text{on-diag}} = \sum_{j=1}^{d_{\text{latent}}} (C_{jj} - \tau)^2$$

where $\tau \in [0, 1]$ is the target correlation. Unlike standard Barlow Twins ($\tau = 1.0$), we use $\tau = 0.3$ to preserve complementarity.

2. Off-diagonal term (decorrelation):

$$(11) \quad \mathcal{L}_{\text{off-diag}} = \sum_{j \neq k} C_{jk}^2$$

This penalizes false correlations and force the model to learn independent features.

The combined Barlow Twins loss is:

$$(12) \quad \mathcal{L}_{\text{BT}} = \mathcal{L}_{\text{on-diag}} + \alpha \mathcal{L}_{\text{off-diag}}$$

where $\alpha = 0.05$ weights the decorrelation term.

Adaptive Multi-Objective Scheduling. To balance contrastive regularization with task-specific supervision, we introduce a time-dependent weighting:

$$(13) \quad \mathcal{L}_{\text{total}}(t) = \mathcal{L}_{\text{CE}} + \lambda(t) \mathcal{L}_{\text{BT}}$$

where \mathcal{L}_{CE} is cross-entropy loss and:

$$(14) \quad \lambda(t) = \lambda_0 \cdot (0.98)^{\lfloor t/5 \rfloor}$$

with $\lambda_0 = 0.01$, t = epoch number.

Early training benefits from strong regularization ($\lambda(t) \approx 0.01$) to establish complementary representations. As the model converges, we progressively emphasize classification ($\lambda(t) \rightarrow 0$), allowing task-specific fine-tuning.

This differs from standard learning rate scheduling (which modulates optimization step size) by dynamically adjusting the objective function itself.

Training Procedure

Given extreme data scarcity ($n=70$ raw samples, 37 for training after split), we apply augmentation with 15× replication:

- Gaussian noise: $\mathbf{x}_{\text{aug}} = \mathbf{x} + \epsilon, \epsilon \sim \mathcal{N}(0, 0.15^2 I)$
- Feature dropout: Randomly zero 30% of features
- Random scaling: Multiply by uniform random factor in [0.7, 1.3]

This expands the effective training set to 555 samples (37×15) while preserving semantic content.

Optimization. We train using AdamW optimizer (Loshchilov and Hutter, 2019) with:

- Learning rate: 5×10^{-4}
- Weight decay: 0.05 (strong L2 regularization)
- Batch size: 8 (limited by small dataset)
- Max epochs: 30

We employ early stopping (patience=5) with learning rate reduction on plateau (factor=0.5, patience=3) to prevent overfitting.

Ensemble Prediction. To mitigate variance from small test set ($n=13$), we train 10 models with different random seeds and combine predictions via weighted ensemble:

$$(15) \quad \hat{y}_{\text{ensemble}} = \sum_{k=1}^{10} w_k \hat{y}_k$$

where weights w_k are proportional to validation accuracy:

$$(16) \quad w_k = \frac{\text{acc}_k^{\text{val}}}{\sum_{j=1}^{10} \text{acc}_j^{\text{val}}}$$

This provides more stable performance estimates than single-seed evaluation.

Experiments

Baseline Architectures

We compare our approach against four state-of-the-art multimodal architectures. For fair comparison, all models use dropout=0.4 and num_layers=1. Latent dimensions follow architectural conventions: latent_dim=64 for Perceiver-based models (standard for cross-attention with our input dimensionality), latent_dim=32 for Transformer/VisualBERT (their typical configurations). This ensures each architecture operates at its conventional capacity relative to model complexity:

Transformer. Naive concatenation-based fusion:

$$(17) \quad \mathbf{x}_{\text{concat}} = [\mathbf{s}; \mathbf{i}] \in \mathbb{R}^{d_s+d_i}$$

followed by standard Transformer encoder with 2 attention heads.

VisualBERT. Pre-trained vision-language model adapted for our sensor-image task. We replace text embeddings with sensor projections while retaining the co-attentional Transformer architecture.

Perceiver. Latent-based architecture using cross-attention from learnable latents to concatenated inputs $[\mathbf{s}; \mathbf{i}]$, followed by 1 self-attention block.

PerceiverIO Classic. Separate Perceiver encoders for each modality (num_latents=4, d_latents=32), with fusion via concatenation of encoded representations followed by MLP decoder. This represents the standard PerceiverIO approach without our modifications.

Ablation Studies

To isolate the contribution of multimodal fusion, we evaluate unimodal baselines:

Sensor-Only: Transformer encoder applied only to \mathbf{s}

Image-Only: Transformer encoder applied only to \mathbf{i}

These ablations share the same architecture (hidden_dim=64, dropout=0.4) as multimodal models for controlled comparison.

Hyperparameter Sensitivity

We conduct a systematic study of the target correlation τ in Equation 10, testing values $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ across 10 seeds each (50 training runs total). This validates that our choice of $\tau = 0.3$ is empirically justified rather than arbitrary.

Evaluation Protocol

Metrics. We report four standard classification metrics:

- Accuracy: Overall correct classification rate
- Weighted F1-score: Harmonic mean of precision/recall, weighted by class frequency
- Weighted Precision: Fraction of correct positive predictions per class
- Weighted Recall: Fraction of actual positives correctly identified per class

All metrics use weighted averaging to account for potential class imbalance.

Statistical Robustness. Each model is trained 10 times with different random seeds. Final predictions uses weighted ensemble (Eqs. 15–16), with performance averaged across all 10 seeds. We report mean values without confidence intervals due to computational constraints, but the consistency of gains across seeds (visible in Figure 5) suggests statistical reliability.

Results

Overall Performance

Table 2 presents performance across all architectures. Our approach achieves 76.9% accuracy and 77.0% weighted F1-score, outperforming all baselines with gains of +25.0% over PerceiverIO Classic, +25.0% over Perceiver, and +43% over Transformer/VisualBERT (Figure 5). Notably,

pre-trained VisualBERT performs identically to standard Transformer (both at 53.8%), confirming that general vision-language representations do not transfer to specialized heritage imaging.

Table 2 – Performance comparison across 10 random seeds. All metrics use weighted averaging.

Model	Accuracy	Weighted F1	Weighted Precision	Weighted Recall
Our Approach	0.769	0.770	0.868	0.769
PerceiverIO Classic	0.615	0.624	0.700	0.615
Perceiver	0.615	0.604	0.670	0.615
VisualBERT	0.538	0.516	0.654	0.538
Transformer	0.538	0.516	0.654	0.538
<i>Unimodal Baselines:</i>				
Sensor Only	0.615	0.591	0.615	0.615
Image Only	0.462	0.405	0.462	0.462

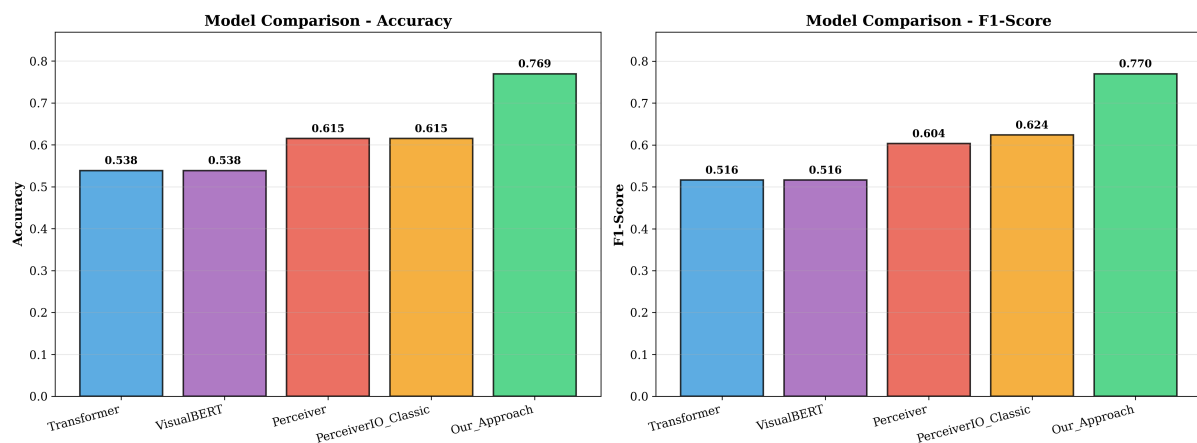


Figure 5 – Model ranking by accuracy (left) and F1-score (right). Our approach achieves 76.9% accuracy, substantially outperforming all baselines

Multimodal Fusion Analysis

Ablation studies reveal that sensor-only achieves 61.5% while image-only reaches 46.2%, demonstrating sensor dominance for degradation assessment (Figure 6, left). Our multimodal fusion achieves 76.9%, representing a +12.5% gain over sensor-only and +50% over image-only (Figure 6, right). This superadditive effect confirms successful complementarity: sensors capture environmental stressors while images reveal visual manifestations that sensors miss. The asymmetric contribution likely reflects that environmental patterns provide more consistent degradation signals than visual inspection alone, particularly in early stages where visual changes are subtle.

Hyperparameter Study

Systematic evaluation of target correlation $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ reveals an unexpected U-shaped performance curve (Figure 7). The optimal value occurs at $\tau = 0.3$ achieving 69.2% accuracy, while extreme decorrelation ($\tau = 0.1$: 53.8%), intermediate values ($\tau \in [0.5, 0.7]$: 53.8%), and strong alignment ($\tau = 0.9$: 61.5%) all show reduced performance. This suggests moderate partial correlation ($\tau = 0.3$) strikes the optimal balance between preserving modality-specific information and maintaining semantic coherence. Our final model trained with $\tau = 0.3$ and refined

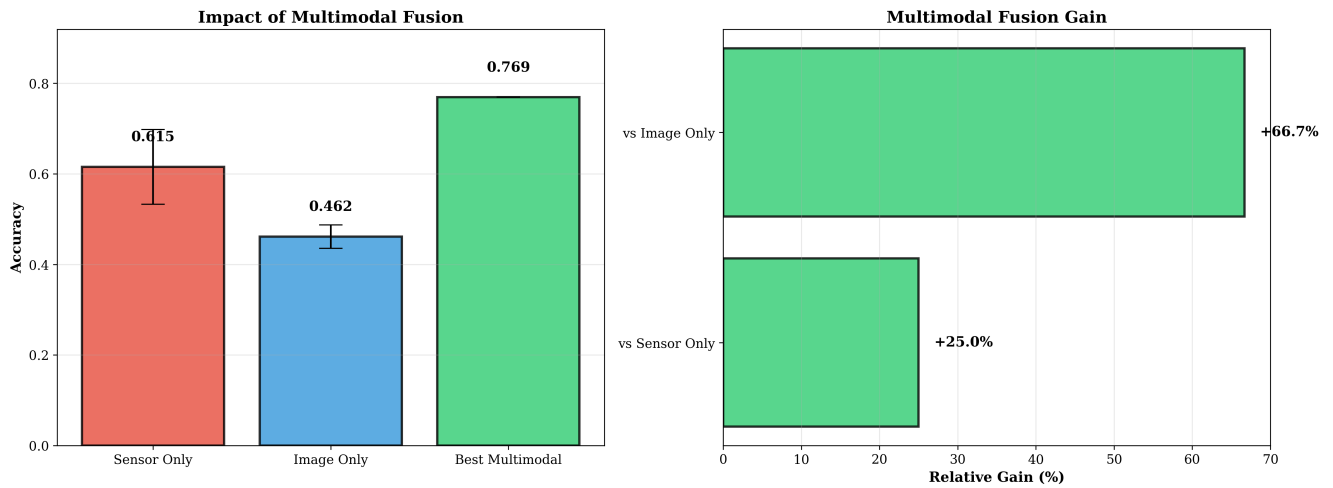


Figure 6 – Multimodal fusion analysis

ensemble strategies achieves 76.9% accuracy through improved regularization techniques. We adopted $\tau = 0.3$ as it empirically demonstrates best modality complementarity.

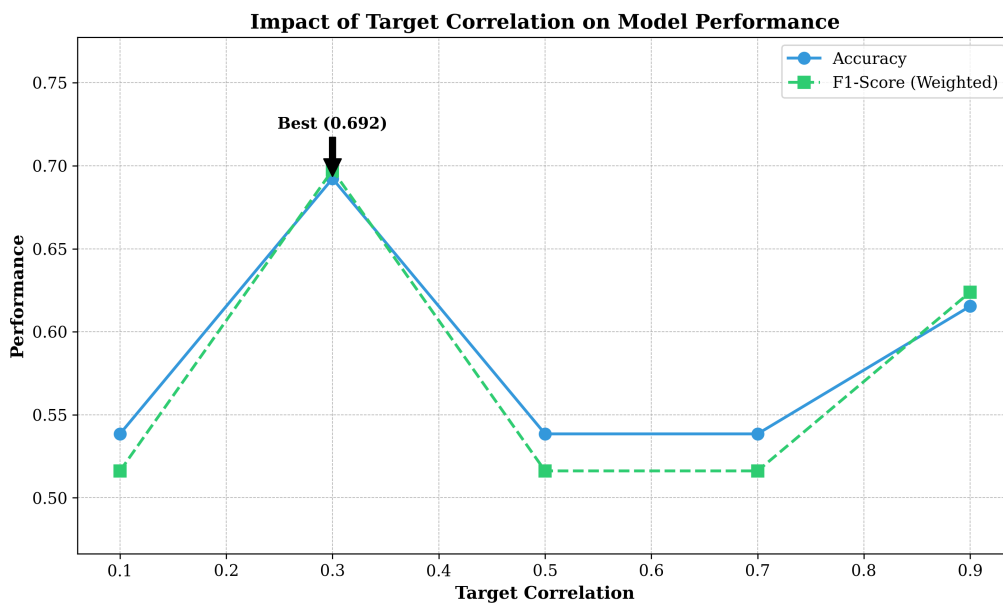


Figure 7 – Impact of target correlation on performance

Error Analysis

The confusion matrix (Figure 8) shows strong performance on Classes 2 and 3 (diagonal entries) with most errors between adjacent degradation levels (Classes 1↔3 and 3↔4), acceptable from a conservation perspective since these distinctions are inherently subtle. No catastrophic misclassifications occur between distant classes, demonstrating coherent severity ordering. Class 0 is absent from the test set due to the stratified split on available data—no blocks with severity level 0 were sampled in the test partition ($n=13$). This reflects the real-world class distribution at the T0 campaign where minimal degradation cases were rare.

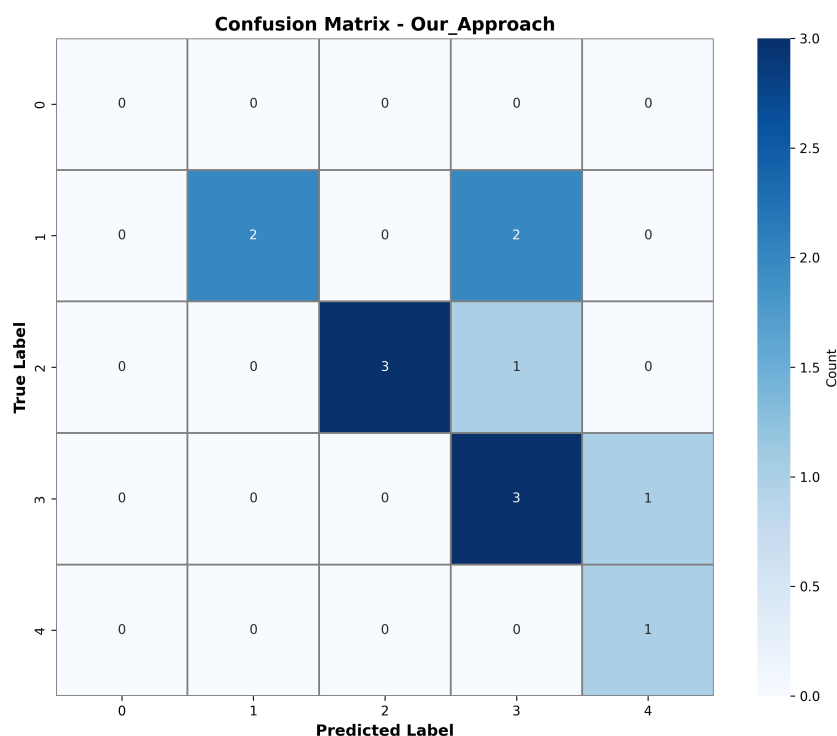


Figure 8 – Confusion matrix showing strong diagonal performance on Classes 2-3

Discussion

Model Performances

Our 76.9% accuracy on 37 training samples (555 after augmentation) represents substantial improvement over architectures designed for large-scale data (+43% vs. Transformer/VisualBERT, +25% vs. PerceiverIO Classic/Perceiver). Two features contributed to this success.

First, adapting Barlow Twins (Zbontar et al., 2021) from view-invariance to modality-complementarity explicitly encourages decorrelation between sensor and image representations. Unlike concatenation-based fusion assuming redundancy, our moderate correlation target ($\tau = 0.3$) preserves modality-specific features while maintaining semantic coherence. The U-shaped hyperparameter curve reveals that both extreme decorrelation ($\tau = 0.1$: 53.8%), intermediate values ($\tau \in [0.5, 0.7]$: 53.8%), and strong alignment ($\tau = 0.9$: 61.5%) underperform the optimal moderate correlation.

Second, architectural simplification through lightweight encoders (12M vs. 50M parameters) prevents overfitting on small datasets. This aligns with sample complexity theory: model capacity should scale with data availability. Pre-trained VisualBERT's failure (53.8%, identical to vanilla Transformer) despite 100K+ training examples confirms that domain shift to scientific heritage imaging renders transfer learning ineffective.

Comparison with Existing Approaches

Cross-attention fusion outperforms concatenation (Transformer: 53.8%) by explicitly modeling inter-modal interactions. However, vanilla Perceiver and PerceiverIO Classic achieve only 61.5%, demonstrating that fusion mechanism alone is insufficient. Our Adaptive Barlow Twins regularization provides a 25.0% gain (from 61.5% to 76.9%) by enforcing complementarity during training.

Few-shot learning methods (Snell et al., 2017) require large meta-training datasets unavailable for heritage monitoring. Our contrastive regularization approach works with a single small dataset, offering an alternative when transfer learning fails due to severe domain shift.

Limitations

Three limitations warrant discussion. First, the small test set ($n=13$) makes each error worth 7.7% accuracy, limiting granular analysis. Our 10-seed ensemble mitigates variance but cannot replace larger evaluation sets. Ongoing campaigns will expand to 200+ samples across three sites by 2026.

Second, results are site-specific to Strasbourg Cathedral with relatively small training data ($n=37$). Generalization across building materials, climates, and degradation mechanisms remains unvalidated. Future work will investigate domain adaptation techniques (Ganin et al., 2016) for cross-site transfer.

Third, the model remains largely black-box. Conservators require interpretability. Integrating Grad-CAM (Selvaraju et al., 2017), Shapley values, and uncertainty quantification would enhance practical deployment.

Conclusion

We presented a lightweight multimodal architecture achieving 76.9% accuracy on small-scale heritage monitoring ($n=37$ train, $n=13$ test), outperforming PerceiverIO/Perceiver by +25.0% and standard baselines by +43%. Three contributions drive this performance: (1) Adaptive Barlow Twins loss encouraging modality complementarity through moderate correlation targets ($\tau = 0.3$, 69.2% accuracy), revealing an optimal balance between alignment and decorrelation superior to extreme values ($\tau = 0.1/0.5/0.7$: 53.8%, $\tau = 0.9$: 61.5%); (2) architectural simplification (12M vs. 50M parameters) preventing overfitting while improving generalization; (3) the ablation study demonstrating superadditive multimodal gains (+25.0% over sensor-only at 61.5%, +66.7% over image-only at 46.2%).

Beyond heritage-specific results, this work shows that contrastive regularization combined with architectural simplicity enables effective multimodal learning when large-scale pre-training is unavailable. Future work will extend to temporal degradation modeling as multi-year data becomes available (2025-2026), integrate explainability techniques for conservator trust, and investigate cross-site transfer learning across diverse heritage contexts.

Acknowledgments

A preprint version of this article has been peer-reviewed and recommended by Peer Community In Archaeo (Hein, 2026) (<https://doi.org/10.24072/pci.archaeo.100708>).

This work was made possible thanks to the funding of the FSP (Fondation des Sciences du Patrimoine) and C2RMF ESPADON. The authors thank the ETIS laboratory (CY Cergy Paris Université / ENSEA / CNRS) for its support, and the company EPITOPOS for providing the data and for their valuable feedback during this initial phase of experimentation.

Funding

This work was funded by the Fondation des Sciences du Patrimoine (FSP).

Conflict of interest disclosure

The authors declare no conflict of interest.

Data, scripts, code and supplementary information availability

All data, scripts, and code used in this study are publicly available on Zenodo (Roqui, 2025) at <https://doi.org/10.5281/zenodo.19053096>. The repository contains: (I) the preprocessed sensor and image features for the Strasbourg Cathedral T0 campaign, (II) the PyTorch implementation of all models (our approach, Transformer, VisualBERT, Perceiver, PerceiverIO Classic, and unimodal baselines), (III) the training and evaluation scripts to reproduce the 10-seed ensemble results reported in this paper, and (IV) a README describing how to install dependencies and run each experiment.

References

- Bengio Y, LeCun Y (1995). Convolutional Networks for Images, Speech, and Time Series. In: *The Handbook of Brain Theory and Neural Networks*. Ed. by Michael A. Arbib. Cambridge, MA: MIT Press, pp. 255–258.
- Cabral FS, Pinto M, Mouzinho F, Fukai H, Tamura S (2020). An Automatic Survey System for Paved and Unpaved Road Classification and Road Anomaly Detection using Smartphone Sensor. arXiv preprint arXiv:2007.13389. <https://doi.org/10.48550/arXiv.2007.13389>. arXiv: 2007.13389 [eess.SP].
- Chen YC, Li L, Yu L, El Kholy A, Ahmed F, Gan Z, Cheng Y, Liu J (2020). UNITER: Universal Image-Text Representation Learning. In: *Proceedings of the European Conference on Computer Vision (ECCV) 2020*. Vol. 12375. Lecture Notes in Computer Science. Cham: Springer, pp. 104–120. https://doi.org/10.1007/978-3-030-58577-8_7.
- Cormier A, Roqui D, Bourges A, Grozavu N (2025). A Comprehensive Study of Weathering Mechanisms and Predictive Modeling for Heritage Stone Conservation. arXiv preprint arXiv:2511.13343. <https://doi.org/10.48550/arXiv.2511.13343>. arXiv: 2511.13343 [cs.CV].
- Dais D, Bal ĪE, Smyrou E, Sarhosis V (2021). Automatic Crack Classification and Segmentation on Masonry Surfaces Using Convolutional Neural Networks and Transfer Learning. *Automation in Construction* **125**, 103606. <https://doi.org/10.1016/j.autcon.2021.103606>.
- Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: *Proceedings of the 9th International Conference on Learning Representations (ICLR)*. <https://doi.org/10.48550/arXiv.2010.11929>. URL: <https://openreview.net/forum?id=YicbFdNTTy>.
- Elgammal A, Liu B, Elhoseiny M, Mazzone M (2017). CAN: Creative Adversarial Networks, Generating “Art” by Learning About Styles and Deviating from Style Norms. Extended version of a paper presented at the 8th International Conference on Computational Creativity (ICCC), Atlanta, GA. <https://doi.org/10.48550/arXiv.1706.07068>. arXiv: 1706.07068 [cs.AI].
- Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, Marchand M, Lempitsky V (2016). Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research* **17**, 1–35. https://doi.org/10.1007/978-3-319-58347-1_10.

- Grilli E, Remondino F (2019). 3D Reconstruction and Semantic Segmentation of Heritage Buildings Using Point Clouds and BIM. In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. Vol. XLII-2/W15. ISPRS, pp. 467–474. <https://doi.org/10.5194/isprs-archives-XLII-2-W15-467-2019>.
- Han S, Pool J, Tran J, Dally WJ (2015). Learning Both Weights and Connections for Efficient Neural Network. In: *Advances in Neural Information Processing Systems 28 (NeurIPS)*, pp. 1135–1143.
- Hein A (2026). A multimodal approach to heritage preservation in the context of climate change. *PCI Archaeology*, 100708. <https://doi.org/10.24072/pci.archaeo.100708>. URL: <https://doi.org/10.24072/pci.archaeo.100708>.
- Hinton G, Vinyals O, Dean J (2015). Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:1503.02531. <https://doi.org/10.48550/arXiv.1503.02531>. arXiv: 1503.02531 [stat.ML].
- Hochreiter S, Schmidhuber J (1997). Long Short-Term Memory. *Neural Computation* 9, 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>.
- Jaegle A, Borgeaud S, Alayrac JB, Doersch C, Ionescu C, Ding D, Koppula S, Zoran D, Brock A, Shelhamer E, Hénaff O, Botvinick MM, Zisserman A, Vinyals O, Carreira J (2022). Perceiver IO: A General Architecture for Structured Inputs and Outputs. In: *Proceedings of the 10th International Conference on Learning Representations (ICLR)*. <https://doi.org/10.48550/arXiv.2107.14795>. URL: <https://openreview.net/forum?id=fILj7WpI-g>.
- Jaegle A, Gimeno F, Brock A, Zisserman A, Vinyals O, Carreira J (2021). Perceiver: General Perception with Iterative Attention. In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 4651–4664.
- Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L (2014). Large-Scale Video Classification with Convolutional Neural Networks. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Columbus, OH, USA: IEEE, pp. 1725–1732. <https://doi.org/10.1109/cvpr.2014.223>.
- Li LH, Yatskar M, Yin D, Hsieh CJ, Chang KW (2019). VisualBERT: A Simple and Performant Baseline for Vision and Language. arXiv preprint arXiv:1908.03557. <https://doi.org/10.48550/arXiv.1908.03557>. arXiv: 1908.03557 [cs.CL].
- Loshchilov I, Hutter F (2019). Decoupled Weight Decay Regularization. In: *Proceedings of the 7th International Conference on Learning Representations (ICLR)*. URL: <https://openreview.net/forum?id=Bkg6RiCqY7>.
- Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY (2011a). Multimodal Deep Learning. In: *Proceedings of the 28th International Conference on Machine Learning (ICML)*. Bellevue, WA, USA, pp. 689–696.
- Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY (2011b). Multimodal Deep Learning. In: *Proceedings of the 28th International Conference on Machine Learning (ICML)*. Bellevue, WA, USA, pp. 689–696.
- Poria S, Cambria E, Bajpai R, Hussain A (2017). A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion. *Information Fusion* 37, 98–125. <https://doi.org/10.1016/j.inffus.2017.02.003>.
- Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I (2021). Learning Transferable Visual Models from Natural Language

- Supervision. In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 8748–8763.
- Roqui D (2025). Code and data for “A Multimodal Approach to Heritage Preservation in the Context of Climate Change”. Zenodo. <https://doi.org/10.5281/zenodo.19053096>. URL: <https://doi.org/10.5281/zenodo.19053096>.
- Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626. <https://doi.org/10.1109/iccv.2017.74>.
- Shalev-Shwartz S, Ben-David S (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press. <https://doi.org/10.1017/CB09781107298019>.
- Singh A, Hu R, Goswami V, Couairon G, Galuba W, Rohrbach M, Kiela D (2022). FLAVA: A Foundational Language and Vision Alignment Model. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15638–15650. <https://doi.org/10.1109/CVPR52688.2022.01519>.
- Snell J, Swersky K, Zemel R (2017). Prototypical Networks for Few-shot Learning. In: *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pp. 4077–4087.
- Turc I, Chang MW, Lee K, Toutanova K (2019). Well-Read Students Learn Better: On the Importance of Pre-training Compact Models. arXiv preprint arXiv:1908.08962v2. <https://doi.org/10.48550/arXiv.1908.08962>. arXiv: 1908.08962 [cs.CL].
- Vaswani A, Bengio S, Brevdo E, Chollet F, Gomez AN, Gouws S, Jones L, Kaiser Ł, Kalchbrenner N, Parmar N, Sepassi R, Shazeer N, Uszkoreit J (2018). Tensor2Tensor for Neural Machine Translation. CoRR abs/1803.07416. <https://doi.org/10.48550/arXiv.1803.07416>. URL: <http://arxiv.org/abs/1803.07416>.
- Vergès-Belmin V, Vallet JM, Bromblet P (2011). Le glossaire illustré ICOMOS-ISCS sur les formes d’altération de la pierre : un outil précieux pour les constats d’état de la statuaire des parcs, jardins et cimetières. In: *Pierre 2011 – Conservation de la pierre dans les parcs, jardins et cimetières, Actes du colloque*. Paris: SFIIC (Section Française de l’Institut International de Conservation).
- Zbontar J, Jing L, Misra I, LeCun Y, Deny S (2021). Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 12310–12320.