STAR-VF: Spatio-Temporal Aware Representation Condensation for Data-Efficient Video Fusion

Yuchen Xian
Zhejiang University
Hangzhou, Zhejiang, China
yc_xian@zju.edu.cn

Yunqiu Xu*
National University of Singapore
Singapore
imyunqiuxu@gmail.com

Guosheng Lin
Nanyang Technological University
Singapore
gslin@ntu.edu.sg

Yang He
CFAR, Agency for Science, Technology and Research
Singapore
he_yang@a-star.edu.sg

Abstract

Video fusion aims to integrate complementary cues from heterogeneous sources, producing visually informative videos with coherent spatio-temporal structures. Jointly learning cross-modal complementarity and temporal dependencies typically relies on large-scale, high-quality video datasets, whose storage and repeated use during optimization incur substantial costs, making data-efficient learning highly desirable. To address this challenge, we rethink video fusion from the perspective of dataset condensation and propose STAR-VF, the first framework that condenses spatio-temporal representations for data-efficient video fusion. STAR-VF condenses the original training set into a compact set of learnable spatio-temporal representations, which are decoded into synthetic video clips and optimized using feature-level supervision provided by a frozen proxy model. The resulting condensed data preserves fusion-relevant spatial and temporal cues without requiring ground-truth fused videos as supervision. To adapt fusion learning to condensed representations, we equip the final fusion model with flow-free inter-frame temporal modeling and query-driven intra-frame aggregation, enabling direct temporal interaction and adaptive selection of fusion-relevant information in representation space. Extensive experiments on infrared-visible, multi-exposure, multi-focus, and medical video fusion benchmarks demonstrate that STAR-VF, trained with only approximately 1/16 of the original temporal training set, achieves fusion quality comparable to methods trained on the full data. Further analyses validate the contributions of learned condensation and flow-free temporal modeling, and demonstrate the robustness of the condensed supervision across different proxy encoders and downstream fusion models. Project page: https://github.com/.

*Corresponding author.

MM '26, Rio de Janeiro, Brazil

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

CCS Concepts

• **Computing methodologies** → **Computer vision tasks**.

Keywords

Video fusion, heterogeneous input, dataset condensation, data-efficient learning, spatio-temporal representation compression

ACM Reference Format:
Yuchen Xian, Yunqiu Xu, Guosheng Lin, and Yang He. 2026. STAR-VF: Spatio-Temporal Aware Representation Condensation for Data-Efficient Video Fusion. In *Proceedings of Proceedings of the 34th ACM International Conference on Multimedia (MM '26).* ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Video fusion [7, 23, 31, 35, 37, 39] is an important yet under-explored task that aims to integrate video sequences from different sensors or vary acquisition conditions into a unified and temporally consistent output, yielding a more informative representation for dynamic scenes. Compared to static image fusion [8, 17, 18, 22, 29–34], video fusion is more challenging, as it must jointly integrate complementary heterogeneous information at the frame level while preserving content stability and temporal structural consistency across frames.

Despite recent advances in video fusion, prior works [7, 23, 37, 39] still follow a full-data training paradigm, learning spatial complementarity and temporal dependency from complete aligned video pairs. Although effective, this paradigm implicitly assumes continuous access to large-scale raw spatiotemporal training data. For video fusion, however, the challenge lies not only in the larger scale and higher training cost of videos, but also in the dual nature of the supervision required for effective fusion: the model must capture complementary information across heterogeneous input sources while preserving temporal coherence over time [7, 11, 37]. Once the training data is compressed, both signals can be degraded simultaneously, weakening the model's ability to learn reliable fusion patterns [12]. Therefore, the central question for dataset compression in video fusion is not whether less data can be used, but **whether the spatio-temporal supervision essential to effective fusion can still be preserved** after substantial data reduction.

Therefore, applying dataset condensation to video fusion **introduces three key challenges**. **(1)** Conventional temporal alignment based on optical flow estimation [2, 24] and feature warping [40] is designed for raw frames or dense features, making it poorly

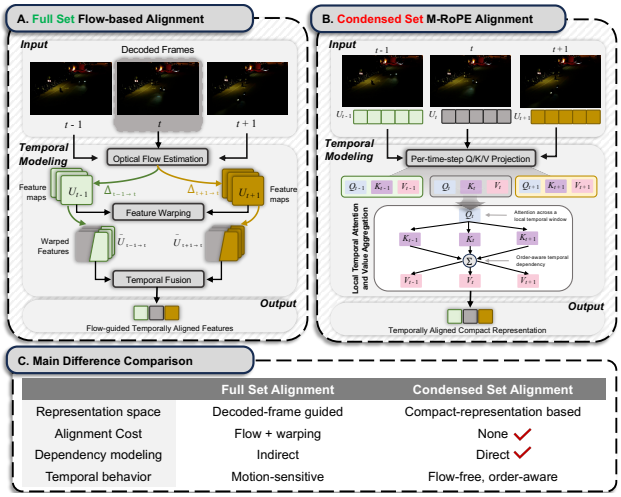

Figure 1: Comparison of temporal alignment paradigms under video fusion dataset condensation. Conventional flow-based alignment is not naturally suited to condensed synthetic clips , whereas our method models order-aware temporal dependency in compact representation space.

matched to compact condensed representations. **(2)** In the condensed setting, fusion-critical cues are no longer expected to be uniformly distributed over all elements, so uniform aggregation can easily dilute the most informative components. **(3)** After substantial data reduction, the condensed training source must still preserve the fusion-relevant supervision required for both cross-source complementarity and temporal coherence. However, these challenges remains largely underexplored. Existing dataset condensation methods are primarily designed for static tasks such as image classification [4, 20]. Therefore, making dataset condensation truly applicable to video fusion requires not only compressing the training data, but also developing a fusion framework that preserves spatio-temporal supervision and remains compatible with condensed representations.

To address this issue, we propose **STAR-VF**, a **S**patio-**T**emporal **A**ware **R**epresentation condensation framework for data-efficient **V**ideo **F**usion. STAR-VF compresses the original training set into a small set of learnable compact spatio-temporal representations, which are decoded into synthetic video clips under feature-level supervision from a frozen proxy network to preserve fusion-relevant signals. Built on this condensed set, the final fusion model is further designed to operate effectively on compact representations, with *flow-free inter-frame temporal modeling* for direct temporal dependency learning in compact space and *query-driven intra-frame aggregation* for adaptive extraction of fusion-critical intra-frame information, shown in Figure 1. Extensive experiments on four representative video fusion tasks show that, using only condensed synthetic data amounting to 1/16 of the original dataset, STAR-VF still achieves performance close to full-data training. These results suggest that spatio-temporal representation condensation can serve as a viable data-efficient learning paradigm for video fusion. The main contributions of this paper are as follows:

(1) We first revisit video fusion from a data-efficient learning perspective and approach it through the lens of dataset condensation, aiming to enable efficient training while retaining fusion-critical spatio-temporal information.
(2) We propose STAR-VF, a new video fusion framework with spatio-temporal aware representation compression, which learns a condensed training source and adapts fusion to condensed representations via flow-free inter-frame temporal modeling and query-driven intra-frame aggregation.
(3) We evaluate STAR-VF on four video fusion tasks, validating that training on condensed data (1/16 of the original dataset) still achieves performance competitive with full-data-trained state-of-the-art approaches.

2 Related Work

Video Fusion. Video fusion extends image fusion from static scenes to dynamic temporal sequences. While existing image fusion works, including CNN- [22, 32], Transformer- [17], GAN- [18], and diffusion-based [33, 38] approaches, have shown strong ability in multi-source information integration, they are primarily designed for single-frame settings and do not explicitly account for temporal dependency [3]. Directly applying image fusion models frame by frame to videos often causes flickering, motion discontinuity, and inter-frame inconsistency. Recent video fusion works [7, 23, 31, 37, 39] therefore place greater emphasis on temporal modeling. UniVF [37] introduces multi-frame learning with flow-guided warping, but suffers from considerable computational overhead. MambaVF [39] alleviates this issue by replacing explicit motion alignment with state-space-based temporal modeling. TemCoCo [7] further improves fused quality and temporal consistency through visual-semantic collaboration. However, these methods still rely on full-data training pipelines and original video representations, leaving compact condensed representations largely unexplored.

Dataset Condensation. Existing dataset condensation methods can be broadly divided into three categories based on their matching objectives . In performance matching, DD [26] represents the meta-based line by learning synthetic samples through bilevel optimization, while RFAD [15] improves kernel-based performance matching with random feature approximation to reduce computational cost. In parameter matching, IDC [9] belongs to the single-step line and improves condensation efficiency through synthetic-data parameterization under limited storage, whereas MTT [1] extends this idea to multi-step parameter matching by aligning training trajectories over multiple optimization steps, thereby alleviating the limitation of single-step matching, which only constrains local updates. In distribution matching, IDM [36] represents the single-layer line and improves naive feature distribution matching by addressing feature-number imbalance and unreliable embeddings for distance computation, while CAFE [25] further advances to multi-layer matching by aligning real and synthetic features across various scales, thus providing stronger global supervision and better discriminative preservation. Although these methods have achieved strong performance, most existing dataset condensation studies are still primarily developed and evaluated on image classification benchmarks [6], leaving limited exploration of multimodal video fusion, where condensed data should preserve not only cross-modal

complementarity but also temporal dependency in compact spatiotemporal representations.

3 Preliminaries

3.1 Problem Formulation

Given a pair of source videos from two modalities, denoted by $(\mathcal{V}_1, \mathcal{V}_2)$, each modality-specific video for $k \in \{1, 2\}$ is written as

$$\mathcal{V}_k = \{\mathbf{I}_{k,t}\}_{t=1}^{T}. \tag{1}$$

Here, $\mathbf{I}_{k,t} \in \mathbb{R}^{C_k \times H \times W}$ denotes the t-th input frame, where C_k is the channel number of modality k, $H \times W$ is the spatial resolution, and T is the full video length.

The goal of multi-source video fusion is to generate a fused video $\widehat{\mathcal{V}}_f = \{\widehat{\mathbf{I}}_{f,t}\}_{t=1}^{T}$ from $(\mathcal{V}_1, \mathcal{V}_2)$. Unlike static image fusion, multi-source video fusion must preserve both *cross-modal complementarity* within each frame and *temporal coherence* across neighboring frames. Accordingly, the task naturally involves two coupled aspects: intra-frame information aggregation and inter-frame temporal dependency modeling. The original training set is denoted by $\mathcal{T}$, which consists of paired source videos.

In data-efficient scenario, the fusion model is not applied to the full sequence at once. Instead, both training and inference are conducted on a local temporal window of length L. Therefore, T denotes the full video length of a sample, whereas L denotes the effective temporal context visible to the model at each step. This distinction is important because the subsequent temporal modeling is window-based rather than full-sequence based.

3.2 Condensed Training Set

Instead of directly training on the full dataset $\mathcal{T}$, we consider a condensed training set $\mathcal{S}$ containing N_s learnable multi-source samples. Each condensed sample is parameterized as a pair of learnable compact spatiotemporal canvases $(\mathbf{G}_1, \mathbf{G}_2)$, where each modality-specific canvas for $k \in \{1, 2\}$ satisfies

$$\mathbf{G}_k \in \mathbb{R}^{C_k \times \rho P_h \times \rho P_w}. \tag{2}$$

Here, ρ denotes the grid size and $P_h \times P_w$ denotes the spatial size of each grid cell. Under this parameterization, one condensed sample contains $T_{\text{cond}} = \rho^2$ compact frame units. To recover video-like inputs from the compact canvases, we introduce a canvas decoder $\mathcal{U}$, which maps each $\mathbf{G}_k$ to a decoded clip:

$$\widetilde{\mathcal{V}}_k = \mathcal{U}(\mathbf{G}_k) = \{\widetilde{\mathbf{I}}_{k,t}\}_{t=1}^{T_{\text{cond}}}. \tag{3}$$

Here, each decoded frame $\widetilde{\mathbf{I}}_{k,t}$ lies in $\mathbb{R}^{C_k \times H \times W}$, matching the channel dimensionality and spatial resolution of the original input frame. These decoded clips serve as the synthetic training source for subsequent fusion learning. Accordingly, $(\mathcal{S}, \mathcal{U})$ defines the condensed-data branch used only during training, while only the final fusion model is retained at test time.

3.3 Challenges under Condensed Representations

Dataset condensation changes not only the amount of training data, but also the representation form on which fusion is learned. Instead of operating on natural video streams directly, the model is trained on decoded clips reconstructed from compact spatiotemporal canvases.

Under this setting, video fusion faces the following three challenges. ***Challenge 1:*** Temporal modeling mismatch. Conventional video fusion often relies on optical flow estimation and feature warping to establish cross-frame correspondence. However, such designs are developed for raw video frames or dense feature maps. In the condensed setting considered here, training is performed on decoded clips reconstructed from compact representations. As a result, conventional flow-based temporal alignment is not naturally matched to the representation form of condensed data. ***Challenge 2:*** Intra-frame information dilution. In condensed representations, fusion-relevant cues are no longer expected to be uniformly distributed over all elements. Instead, they are more likely to concentrate in a sparse informative subset. Consequently, uniform aggregation can dilute the most useful information for fusion, making it difficult for the model to emphasize the elements that are truly critical. ***Challenge 3:*** Supervision preservation under data compression. Effective video fusion depends on two coupled types of supervision: cross-modal complementarity within each frame and temporal coherence across neighboring frames. Once the training set is compressed into a small number of synthetic samples, both signals may be weakened simultaneously. Therefore, the last challenge is to preserve the fusion-relevant spatio-temporal supervision after substantial data reduction.

These three challenges show that dataset condensation for multi-source video fusion is not merely a data reduction problem. It also requires a fusion pipeline that remains compatible with condensed spatio-temporal representations.

4 STAR-VF

4.1 Overview

As shown in Figure 2, STAR-VF contains a condensed-data branch and a final fusion model. The condensed set is

$$\mathcal{S} = \{(\mathbf{G}_{1,s}, \mathbf{G}_{2,s})\}_{s=1}^{N_s}, \qquad \mathbf{G}_{k,s} \in \mathbb{R}^{C_k \times \rho P_h \times \rho P_w}, \tag{4}$$

where $k \in \{1, 2\}$. A modality-conditioned decoder $\mathcal{U}$ maps the $\rho \times \rho$ cells of each canvas, in row-major order $u = (a-1)\rho + b$, to $\widetilde{\mathcal{V}}_{k,s} = \mathcal{U}(\mathbf{G}_{k,s}; k) = \{\widetilde{\mathbf{I}}_{k,s,u}\}_{u=1}^{\rho^2}$. Let N_{real} be the number of paired temporal units in the original set. The training-source ratio is $r_{\text{cond}} = \frac{N_s \rho^2}{N_{\text{real}}}$. The canvases are optimized variables rather than selected clips, and $\mathcal{U}$ converts their cells into full-resolution training inputs. We use $\rho = 4$ and choose N_s such that $r_{\text{cond}} = 1/16$. This ratio measures temporal training units rather than spatial downsampling.

Training has two stages. Stage I optimizes $\mathcal{S}$ and $\mathcal{U}$ by matching statistics extracted by a frozen proxy. Stage II fixes the condensed branch and trains $\mathcal{F}_\Theta$ on paired synthetic clips. For each target time t, the final model encodes a local window, models inter-frame dependency, aggregates informative elements from both modalities, and reconstructs the fused center frame. At inference, only $\mathcal{F}_\Theta$ is retained to process real video pairs.

4.2 Flow-Free Inter-Frame Temporal Modeling

Decoded synthetic frames do not provide reliable motion fields for feature warping. We therefore model temporal dependency directly

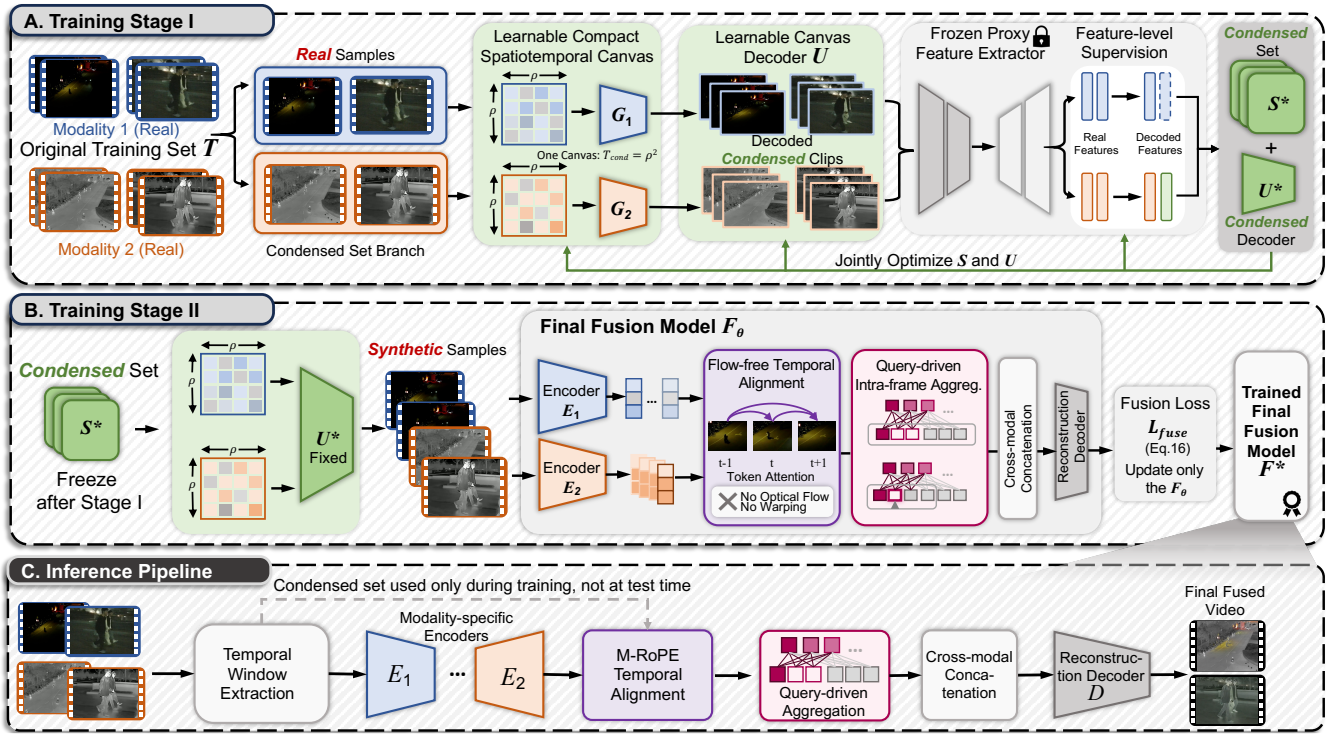

Figure 2: Overview of the proposed STAR-VF framework. In Stage I, learnable compact spatio-temporal canvases and the canvas decoder U are jointly optimized under feature-level supervision from a frozen proxy feature extractor, yielding the learned condensed set S^* and decoder U^*. In Stage II, the learned condensed set is decoded into synthetic temporal samples to train the final fusion model F_θ with flow-free temporal alignment and query-driven intra-frame aggregation. During inference, only the trained fusion model is used, without the condensed set or proxy model.

in representation space, allowing neighboring frames to interact without optical-flow estimation. This module uses the center frame as the query and the complete local window as temporal memory.

For a neighborhood $\mathcal{N}(t)$ of length L, the modality-specific encoder produces

$$\mathbf{X}_{k,\tau} = \mathcal{E}_k(\mathbf{I}_{k,\tau}) \in \mathbb{R}^{J\times d}, \qquad \tau \in \mathcal{N}(t), k \in \{1,2\}, \tag{5}$$

where J is the number of tokens. Let h be the number of heads and $d_h = d/h$. For head m,

$$\mathbf{Q}_{k,t}^{(m)} = \mathbf{X}_{k,t}\mathbf{W}_Q^{(m)}, \quad \mathbf{K}_{k,\tau}^{(m)} = \mathbf{X}_{k,\tau}\mathbf{W}_K^{(m)}, \quad \mathbf{V}_{k,\tau}^{(m)} = \mathbf{X}_{k,\tau}\mathbf{W}_V^{(m)}, \tag{6}$$

with $\mathbf{W}_Q^{(m)}, \mathbf{W}_K^{(m)}, \mathbf{W}_V^{(m)} \in \mathbb{R}^{d\times d_h}$. For token j, M-RoPE uses $\mathbf{p}_{\tau,j} = (\tau, x_j, y_j)$ and $\Psi(\mathbf{v}, \mathbf{p}_{\tau,j}) = \mathbf{R}_t(\tau)\mathbf{R}_x(x_j)\mathbf{R}_y(y_j)\mathbf{v}$ to encode temporal order and spatial position. We concatenate the encoded keys and values over the window:

$$\overline{\mathbf{K}}_{k,t}^{(m)} = \underset{\tau\in\mathcal{N}(t)}{\text{Concat}}\, \Psi(\mathbf{K}_{k,\tau}^{(m)}, \mathbf{p}_\tau), \quad \overline{\mathbf{V}}_{k,t}^{(m)} = \underset{\tau\in\mathcal{N}(t)}{\text{Concat}}\, \mathbf{V}_{k,\tau}^{(m)}, \tag{7}$$

both in $\mathbb{R}^{LJ\times d_h}$. Temporal attention is normalized over all LJ key tokens:

$$\mathbf{A}_{k,t}^{(m)} = \text{Softmax}_{LJ}\left(\frac{\Psi(\mathbf{Q}_{k,t}^{(m)}, \mathbf{p}_t)\overline{\mathbf{K}}_{k,t}^{(m)\top}}{\sqrt{d_h}}\right), \qquad \mathbf{H}_{k,t}^{(m)} = \mathbf{A}_{k,t}^{(m)}\overline{\mathbf{V}}_{k,t}^{(m)}. \tag{8}$$

Joint normalization lets tokens from different frames compete while keeping the response scale independent of L. M-RoPE retains their temporal order and spatial location.

The enhanced center-frame representation is

$$\mathbf{Z}_{k,t} = \text{Proj}\left[\text{Concat}(\mathbf{H}_{k,t}^{(1)}, \ldots, \mathbf{H}_{k,t}^{(h)})\right] \in \mathbb{R}^{J\times d}. \tag{9}$$

4.3 Query-Driven Intra-Frame Aggregation

Fusion-relevant cues are not distributed uniformly across the J tokens, so direct pooling may suppress localized or modality-specific details. For each modality, M learnable queries $\mathbf{Q}_k^{\text{agg}} \in \mathbb{R}^{M\times d}$ summarize the enhanced representation into aggregation slots. With $\mathbf{W}_Q^k, \mathbf{W}_K^k, \mathbf{W}_V^k \in \mathbb{R}^{d\times d}$, we compute

$$\widehat{\mathbf{Q}}_k = \mathbf{Q}_k^{\text{agg}}\mathbf{W}_Q^k, \qquad \widehat{\mathbf{K}}_{k,t} = \mathbf{Z}_{k,t}\mathbf{W}_K^k, \qquad \widehat{\mathbf{V}}_{k,t} = \mathbf{Z}_{k,t}\mathbf{W}_V^k, \tag{10}$$

$$\mathbf{Y}_{k,t} = \text{Softmax}_J\left(\frac{\widehat{\mathbf{Q}}_k\widehat{\mathbf{K}}_{k,t}^\top}{\sqrt{d}}\right)\widehat{\mathbf{V}}_{k,t} \in \mathbb{R}^{M\times d}. \tag{11}$$

The row-wise attention weights allow each slot to select a different subset of compact elements. This produces a structured summary for each modality while retaining non-uniform information selection.

The modality features are concatenated along the feature dimension and reconstructed by $\mathcal{D} : \mathbb{R}^{M\times 2d} \to \mathbb{R}^{C_f\times H\times W}$:

$$\widehat{\mathbf{I}}_{f,t} = \mathcal{D}\left[\text{Concat}_d(\mathbf{Y}_{1,t}, \mathbf{Y}_{2,t})\right]. \tag{12}$$

Table 1: Quantitative comparison on IVF (VTMOT), MVF (Harvard), MEF (YouTube-HDR), and MFF (DAVIS). The training-data ratio r_{train} denotes the fraction of the original temporal training units used to train each method. STAR-VF uses only approximately 1/16 of the original training source, yet achieves fusion quality comparable to the full-data baselines. Red and Blue indicate the best and second-best metric values.

Method	Train Ratio	IVF (VTMOT [41])					MVF (Harvard [21])					MEF (YouTube-HDR [27])					MFF (DAVIS [19])				
		CC↑	EN↑	SSIM↑	BiSWE↓	MS2R↓	CC↑	EN↑	SSIM↑	BiSWE↓	MS2R↓	CC↑	EN↑	SSIM↑	BiSWE↓	MS2R↓	CC↑	EN↑	SSIM↑	BiSWE↓	MS2R↓
UniVF [37]	1	0.65	6.81	0.65	3.95	0.35	0.78	4.27	0.75	29.92	1.23	0.96	7.25	0.65	6.44	0.34	0.98	7.41	0.90	6.04	1.11
VideoFusion [23]	1	0.62	6.59	0.62	3.14	0.59	0.78	4.91	0.69	19.71	1.31	0.91	6.82	0.58	9.98	1.82	0.96	7.34	0.82	9.07	6.13
TemCoCo [7]	1	0.63	6.68	0.60	6.26	0.59	0.77	5.40	0.24	28.41	1.57	0.90	7.34	0.53	11.62	1.70	0.97	7.45	0.84	12.34	4.08
MambaVF [39]	1	-	-	0.65	3.91	0.36	-	-	0.76	29.61	1.27	-	-	-	6.96	0.16	-	-	-	8.29	-
STAR-VF (ours)	∼1/16	0.64	6.73	0.65	3.91	0.35	0.80	4.37	0.76	28.36	1.26	0.91	7.28	0.55	10.23	1.28	0.96	7.42	0.91	5.53	1.18

Algorithm 1 Two-Stage Training Pipeline for Video Fusion

Require: training set $\mathcal{T}$, condensed set size N_s, and window length L
Ensure: $(\mathcal{S}^*, \mathcal{U}^*, \mathcal{F}_\Theta^*)$

Initialize $\mathcal{S}$ and $\mathcal{U}$
Stage I: Condensation
for each iteration **do**
  Sample a real clip pair from $\mathcal{T}$ ▷ real supervision source
  Sample $s \in \{1, \dots, N_s\}$ and decode $\widetilde{\mathcal{V}}_{k,s} \leftarrow \mathcal{U}(\mathbf{G}_{k,s}; k)$ for $k \in \{1, 2\}$ ▷ synthetic clips from compact canvases
  Compute $\mathcal{L}_{\text{cond}}$ and update $\mathcal{S}, \mathcal{U}$ ▷ learn condensed branch
end for
$(\mathcal{S}^*, \mathcal{U}^*) \leftarrow (\mathcal{S}, \mathcal{U})$ ▷ freeze for Stage II

Initialize $\mathcal{F}_\Theta$
Stage II: Fusion Training
for each iteration **do**
  Decode synthetic clips from $(\mathcal{S}^*, \mathcal{U}^*)$
  Construct $\mathcal{W}_t$ with length L ▷ local temporal window
  Predict $\widehat{\mathbf{I}}_{f,t} \leftarrow \mathcal{F}_\Theta(\mathcal{W}_t)$ ▷ center-frame fusion
  Compute $\mathcal{L}_{\text{fuse}}$ and update Θ ▷ train final fusion model
end for
$\mathcal{F}_\Theta^* \leftarrow \mathcal{F}_\Theta$
return $(\mathcal{S}^*, \mathcal{U}^*, \mathcal{F}_\Theta^*)$

4.4 Condensed Set Learning and Final Fusion Training

Stage I: Condensed Set Learning. For each sampled real clip and decoded synthetic clip, a frozen proxy $\phi_{k,\ell}$ [37] extracts multi-level features

$$\mathbf{F}^r_{k,\ell,n,t} = \phi_{k,\ell}(\mathbf{I}^r_{k,n,t}), \qquad \mathbf{F}^s_{k,\ell,s,u} = \phi_{k,\ell}(\widetilde{\mathbf{I}}_{k,s,u}). \tag{13}$$

Following multi-level feature matching [25], we use

$$\mathcal{L}_{\text{cond}} = \underbrace{\mathcal{D}^{16}_{\text{mse}} + \lambda_1 \mathcal{D}_{\text{ch}} + \lambda_2 \left(\mathcal{D}^{4}_{\text{sp}} + \mathcal{D}^{8}_{\text{sp}} \right)}_{\text{multi-level feature matching}} + \underbrace{\lambda_3 \mathcal{L}_{\text{color}}}_{\text{pixel-space matching}} . \tag{14}$$

Here $\mathcal{D}^{16}_{\text{mse}}$ matches mean 16×16 proxy feature maps, $\mathcal{D}_{\text{ch}}$ matches channel mean and standard deviation, and $\mathcal{D}^{q}_{\text{sp}}$ matches spatial mean and standard deviation on a $q \times q$ grid. Each term is summed over modalities and proxy levels after pooling over sampled clips and frames. $\mathcal{L}_{\text{color}}$ applies the same moment matching in pixel space [5, 14], preventing feature matching from drifting toward implausible image statistics. Stage I therefore preserves representative spatial statistics in a compact paired source.

Stage II: Final Fusion Model Training. After fixing $(\mathcal{S}, \mathcal{U})$, we train $\mathcal{F}_\Theta$ from scratch on local windows $\mathcal{W}_t$ sampled from the decoded clips:

$$\widehat{\mathbf{I}}_{f,t} = \mathcal{F}_\Theta(\mathcal{W}_t). \tag{15}$$

The fusion objective is

$$\mathcal{L}_{\text{fuse}} = \mathcal{L}_{\text{int}} + \mathcal{L}_{\text{grad}} + \mathcal{L}_{\text{ssim}} + \lambda_4 \mathcal{L}_{\text{temp}}. \tag{16}$$

We use the intensity and gradient objectives of [32], SSIM [28], and the local temporal-consistency objective of [37] without modification. The first three terms supervise complementary source content in each fused frame, while $\mathcal{L}_{\text{temp}}$ regularizes changes across adjacent predictions.

5 Experiments

5.1 Experimental Setup

Datasets and Metrics. We evaluate STAR-VF on four video-fusion tasks: infrared-visible fusion (IVF) on VTMOT [41], medical video fusion (MVF) on Harvard [21], multi-exposure fusion (MEF) on YouTube-HDR [27], and multi-focus fusion (MFF) on DAVIS [19]. These datasets are used under their corresponding fusion tasks following the UniVF protocol, without additional custom pairing, alignment, or split construction. We compare with the dedicated video-fusion methods UniVF [37], VideoFusion [23], TemCoCo [7], and MambaVF [39].

We report the image-level metrics CC, EN, and SSIM used in IVIF-ZOO [13], together with the video-level temporal metrics BiSWE and MS2R from VF-Bench [37]. Higher CC, EN, and SSIM and lower BiSWE and MS2R indicate better performance. Efficiency is measured by parameters, FLOPs, latency, and peak GPU memory.

Implementation Details. In Stage I, compact canvases are initialized by sampling and spatially interpolating real frames. A frozen UniVF model provides proxy features, and the condensed branch is optimized for 10,000 iterations using Adam [10] with $(\beta_1, \beta_2) = (0.9, 0.999)$ and a learning rate of 10^{-2}. In Stage II, the condensed branch is fixed and a fusion model with $C = 32$ and four Transformer blocks is trained from scratch for 20,000 iterations. We use $M = 8$ aggregation queries and the centered 3-frame window $\mathcal{N}(t) = \{t-1, t, t+1\}$. AdamW [16], a cosine schedule from $2{\times}10^{-4}$ to 10^{-6}, and a batch size of 8 are used. We empirically set λ_1, λ_2, λ_3 and λ_4 to 0.5, 0.1, 0.1 and 0.5, respectively.

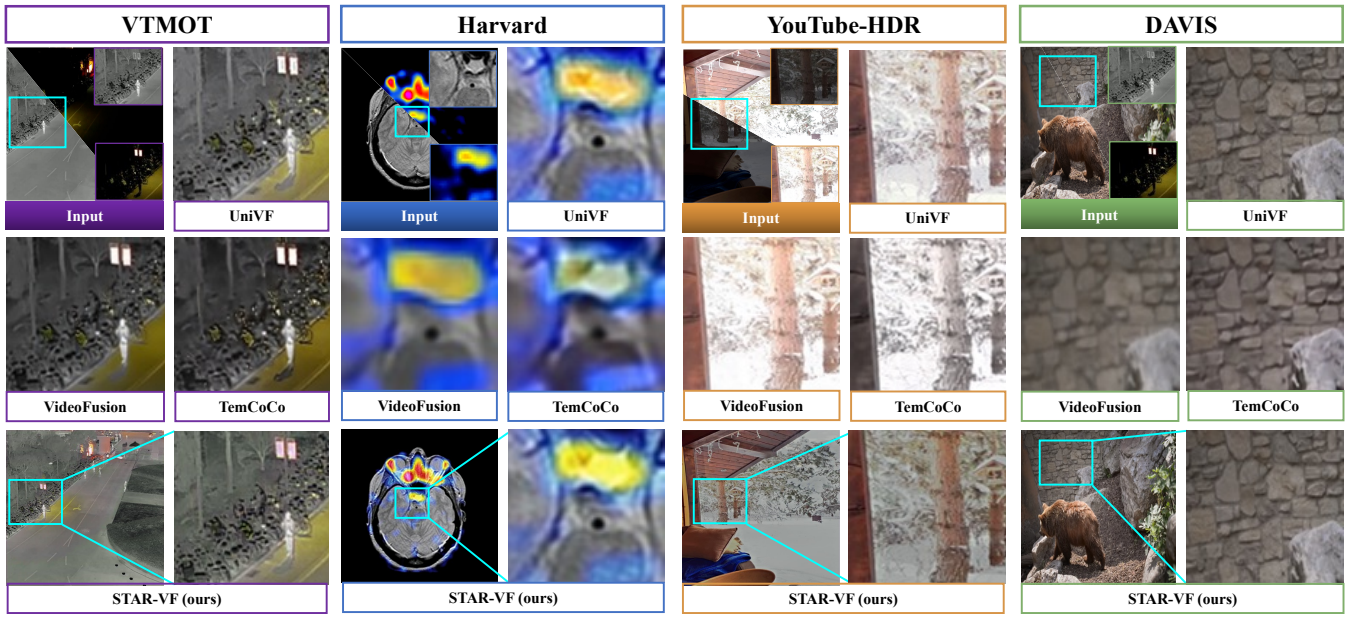

Figure 3: Qualitative comparison on four representative benchmarks. For each example, we show the input pair, our method, the full-size variant trained on the original data, and competing video fusion baselines (*VideoFusion*, *UniVF*, and *TemCoCo*). The cyan boxes indicate enlarged local regions for detailed comparison.

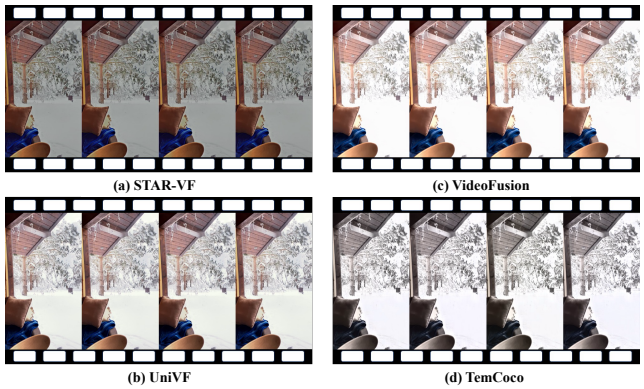

Figure 4: Frame-by-frame qualitative comparison on a representative multi-source video fusion sequence.

Table 2: Training-source budgets used by STAR-VF.

Dataset	Task	Train seqs	N_{real}	N_s	N_{cond}	r_{cond}
VTMOT	IVF	75	22,500	88	1,408	1/16.0
Harvard	MVF	49	1,308	6	96	1/13.6
YouTube-HDR	MEF	450	66,595	261	4,176	1/15.9
DAVIS	MFF	120	8,195	32	512	1/16.0

Table 3: Computational efficiency comparison.

Model	Resolution	Params	GFLOPs	Latency	Memory
UniVF [37]	256×256	9.16M	934.7	531.4ms	696.8MB
Condensed (ours)	256×256	0.32M	184.0	110.5ms	696.8MB
UniVF [37]	480×854	9.16M	6139.1	820.7ms	4110.5MB
Condensed (ours)	480×854	0.32M	1066.6	468.5ms	4110.5MB

5.2 Main Results

Quantitative Results. Table 1 shows that STAR-VF, trained with the dataset-specific synthetic budgets in Table 2, remains competitive with full-data methods. It ranks first in CC and SSIM on MVF and in SSIM and BiSWE on MFF, while remaining close to the best MS2R, showing that a compact training source retains useful per-frame fusion cues and temporal regularity.

Qualitative Results. Figure 3 shows that STAR-VF preserves salient structures and complementary details across all four tasks, including pedestrian contours on VTMOT, tissue boundaries on Harvard, highlight and shadow details on YouTube-HDR, and focus transitions on DAVIS. The enlarged regions expose local structural and brightness differences. In Figure 4, STAR-VF also exhibits fewer brightness fluctuations and more consistent colors than the compared methods.

Training-Source Efficiency. Table 2 reports the actual budget on each benchmark, where $N_{\text{cond}} = N_s T_{\text{cond}}$. All settings use $\rho = 4$, $P_h \times P_w = 64 \times 64$, and $T_{\text{cond}} = 16$. STAR-VF reduces the training source to approximately 1/16 of the original temporal units on three datasets and 1/13.6 on Harvard, substantially lowering the data required for fusion training.

Computational Efficiency. Table 3 compares the deployed fusion architectures; the condensed branch and proxy are absent at inference. Relative to UniVF, STAR-VF reduces parameters from 9.16M to 0.32M. At 256×256, FLOPs and latency decrease from 934.7G and 531.4 ms to 184.0G and 110.5 ms; at 480×854, they decrease from 6139G and 820.7 ms to 1066.6G and 468.5 ms. These inference gains arise from the final lightweight architecture and are distinct from the reduction in training-data budget.

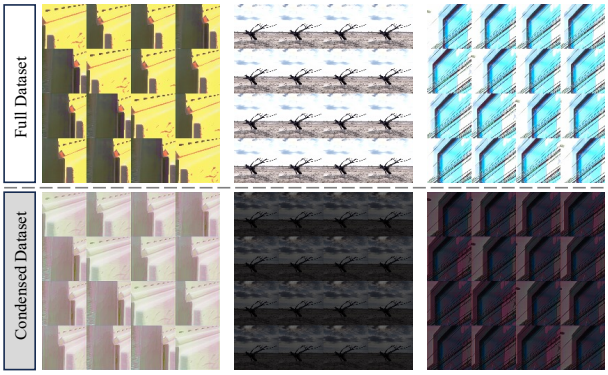

Figure 5: Visual comparison between the full dataset and the learned condensed dataset.

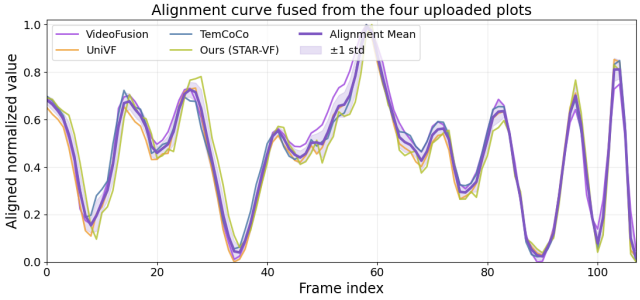

Figure 6: Alignment consistency comparison across video fusion methods, evaluated by frame-wise aligned ROI luminance trajectories after min-max normalization

Frame-Wise Consistency. We extract luminance trajectories from aligned regions of interest and apply min–max normalization to remove scale differences. Figure 6 shows that STAR-VF follows the shared temporal trend with smaller local deviations, indicating a more stable fused appearance.

5.3 Ablation Studies

Effect of Flow-Free Temporal Modeling. Using the complete training set, we replace VideoFusion's RAFT-based alignment and feature warping with M-RoPE attention while retaining the same fusion setting. Table 4 shows comparable fusion quality and temporal consistency on IVF and MFF, indicating that direct representation-level interaction removes explicit flow estimation without sacrificing performance. Because condensation is disabled in this comparison, the result isolates the contribution of the temporal module from the effect of reducing the training source.

Effect of Condensation Grid Size. Table 5 reveals a non-monotonic trade-off. The 2×2, 3×3, and 5×5 grids favor image-level CC/SSIM but yield BiSWE values of approximately 5.07–5.17, indicating stronger temporal fluctuation. The 4×4 grid does not maximize per-frame correlation, but achieves the lowest MS2R (0.35) and lower BiSWE (3.91). We therefore use 4×4 as the best balance between representation capacity, compactness, and temporal stability.

Attribution to Learned Condensation. On VTMOT, using the same final architecture and 1/16 budget, five random compact sources obtain 5.64±0.79 EN, 0.48±0.12 CC, and 0.51±0.08 SSIM, whereas learned condensation achieves 6.73, 0.64, and 0.65. This controlled comparison attributes the gain to optimized condensation rather than model size or arbitrary data reduction. The substantial gap under an identical budget further shows that the number of synthetic frames alone cannot explain the observed performance.

Table 4: Ablation of flow-free temporal modeling on IVF.

Method	CC↑	EN↑	SSIM↑	BiSWE↓	MS2R↓
UniVF [37]	0.65	6.81	0.65	3.95	0.35
w/o condensation	0.64	6.73	0.64	3.91	0.36
Ours (w/o RAFT)	0.64	6.73	0.65	3.91	0.35

Table 5: Ablation on condensation grid size on IVF.

Grid	#Frames	CC↑	EN↑	SSIM↑	BiSWE↓	MS2R↓
1×1 (full)	1	0.64	6.73	0.64	3.91	0.36
2×2	4	0.90	7.25	0.87	5.07	0.36
3×3	9	0.89	7.82	0.88	5.10	0.36
4×4 (ours)	16	**0.64**	**6.73**	**0.65**	**3.91**	**0.35**
5×5	25	0.90	6.62	0.88	5.17	0.36

Table 6: Proxy robustness and downstream transferability on VTMOT. "Full" denotes UniVF trained on the complete real set; all other rows use a 1/16 condensed source.

Downstream	Source/Proxy	EN↑	CC↑	SSIM↑	BiSWE↓	MS2R↓
UniVF	Full	6.814	0.654	0.652	3.950	0.353
UniVF	Default	6.370	0.625	0.626	5.505	0.379
UniVF	ResNet-18	6.320	0.601	0.626	5.101	0.382
UniVF	VGG-16	6.510	0.660	0.628	5.205	0.375
STAR-VF	Default	6.731	0.641	0.648	3.920	0.356
STAR-VF	ResNet-18	6.720	0.591	0.635	3.586	0.355
STAR-VF	VGG-16	6.709	0.643	0.646	3.445	0.354

Proxy Robustness and Transferability. Table 6 shows that ResNet-18 and VGG-16 proxies produce condensed sources comparable to the default encoder for STAR-VF. The same sources also train UniVF effectively, with the VGG-16 proxy reaching 6.510 EN, 0.660 CC, and 0.628 SSIM. Thus, the condensed supervision is not tied to a particular proxy or downstream fusion model.

6 Conclusion

In this paper, we introduced STAR-VF, the first spatio-temporal representation condensation framework for data-efficient video fusion. Across four representative tasks, STAR-VF achieves performance comparable to full-data-trained methods using only a 1/16 training-source budget, demonstrating that compact synthetic representations can preserve fusion-relevant spatial and temporal cues. Controlled comparisons further show that the improvement arises from learned condensation rather than random compact construction or model-size reduction. Experiments with different proxy networks also indicate that the condensed supervision is transferable and not tied to a specific feature extractor or downstream model. Since the condensed branch and proxy are used only during training, inference operates directly on real multi-source videos using the final fusion model. Nevertheless, performance remains bounded by the capacity of the condensed source, and a fixed budget may become insufficient as the diversity of the original data increases.

Acknowledgments

This research is supported by the RIE2025 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) (Award I2301E0026), administered by A*STAR, as well as supported by Alibaba Group and NTU Singapore through Alibaba-NTU Global e-Sustainability CorpLab (ANGEL).

References

[1] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. 2022. Dataset distillation by matching training trajectories. In *CVPR Workshop*.

[2] Kelvin C.K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. 2022. BasicVSR++: Improving Video Super-Resolution With Enhanced Propagation and Alignment. In *CVPR*.

[3] Guo Chen, Yifei Huang, Jilan Xu, Baoqi Pei, Jiahao Wang, Zhe Chen, Zhiqi Li, Tong Lu, and Limin Wang. 2026. Video Mamba Suite: State Space Model as a Versatile Alternative for Video Understanding. *IJCV* (2026).

[4] Jianrong Ding, Zhanyu Liu, Guanjie Zheng, Haiming Jin, and Linghe Kong. 2024. CondTSF: One-line Plugin of Dataset Condensation for Time Series Forecasting. In *NeurIPS*.

[5] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *CVPR*.

[6] Jiahui Geng, Zongxiong Chen, Yuandou Wang, Herbert Woisetschlaeger, Sonja Schimmler, Ruben Mayer, Zhiming Zhao, and Chunming Rong. 2023. A survey on dataset distillation: Approaches, applications and future directions. In *IJCAI*.

[7] Meiqi Gong, Hao Zhang, Xunpeng Yi, Linfeng Tang, and Jiayi Ma. 2025. Temcoco: Temporally consistent multi-modal video fusion with visual-semantic collaboration. In *ICCV*.

[8] Dingli Hua, Qingmao Chen, Zhiliang Wu, Yifan Zuo, Wenying Wen, and Yuming Fang. 2025. Perceptual Transform Fusion of Infrared and Visible Images. *IEEE TCSVT* (2025).

[9] Jang-Hyun Kim, Jinuk Kim, Seong Joon Oh, Sangdoo Yun, Hwanjun Song, Joonhyun Jeong, Jung-Woo Ha, and Hyun Oh Song. 2022. Dataset condensation via efficient synthetic-data parameterization. In *ICML*.

[10] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

[11] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. 2024. VideoMamba: State Space Model for Efficient Video Understanding. In *ECCV*.

[12] Zhe Li, Hadrien Reynaud, and Bernhard Kainz. 2025. Leveraging Multi-Modal Information to Enhance Dataset Distillation. In *BMVC Workshop*.

[13] Jinyuan Liu, Guanyao Wu, Zhu Liu, Di Wang, Zhiying Jiang, Long Ma, Wei Zhong, Xin Fan, and Risheng Liu. 2024. Infrared and visible image fusion: From data compatibility to task adaption. *IEEE TPAMI* (2024).

[14] Jinyuan Liu, Bowei Zhang, Qingyun Mei, Xingyuan Li, Yang Zou, Zhiying Jiang, Long Ma, Risheng Liu, and Xin Fan. 2025. Dcevo: Discriminative cross-dimensional evolutionary learning for infrared and visible image fusion. In *CVPR*.

[15] Noel Loo, Ramin Hasani, Alexander Amini, and Daniela Rus. 2022. Efficient dataset distillation using random feature approximation. In *NeurIPS*.

[16] Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

[17] Jiayi Ma, Linfeng Tang, Fan Fan, Jun Huang, Xiaoguang Mei, and Yong Ma. 2022. SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica* (2022).

[18] Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, and Junjun Jiang. 2019. FusionGAN: A generative adversarial network for infrared and visible image fusion. *Information Fusion* (2019).

[19] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. 2017. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675* (2017).

[20] Ding Qi, Jian Li, Junyao Gao, Shuguang Dou, Ying Tai, Jianlong Hu, Bo Zhao, Yabiao Wang, Chengjie Wang, and Cairong Zhao. 2025. Towards Universal Dataset Distillation via Task-Driven Diffusion. In *CVPR*.

[21] D Summers. 2003. Harvard Whole Brain Atlas: www. med. harvard. edu/AANLIB/home. html. *Journal of Neurology, Neurosurgery & Psychiatry* (2003).

[22] Han Tang, Bin Xiao, Weisheng Li, and Guoyin Wang. 2018. Pixel Convolutional Neural Network for Multi-Focus Image Fusion. *Information Sciences* (2018).

[23] Linfeng Tang, Yeda Wang, Meiqi Gong, Zizhuo Li, Yuxin Deng, Xunpeng Yi, Chunyu Li, Han Xu, Hao Zhang, and Jiayi Ma. 2026. VideoFusion: A Spatio-Temporal Collaborative Network for Multi-modal Video Fusion. In *CVPR*.

[24] Zachary Teed and Jia Deng. 2020. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In *ECCV*.

[25] Kai Wang, Bo Zhao, Xiangyu Peng, Zheng Zhu, Shuo Yang, Shuo Wang, Guan Huang, Hakan Bilen, Xinchao Wang, and Yang You. 2022. CAFE: Learning to Condense Dataset by Aligning Features. In *CVPR*.

[26] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. 2018. Dataset Distillation. *arXiv preprint arXiv:1811.10959* (2018).

[27] Yilin Wang, Joong Gon Yim, Neil Birkbeck, and Balu Adsumilli. 2024. YouTube SFV+HDR Quality Dataset. In *ICIP*.

[28] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image Quality Assessment: from Error Visibility to Structural Similarity. *IEEE TIP* (2004).

[29] Zhiliang Wu, Kang Zhang, Hanyu Xuan, Xia Yuan, and Chunxia Zhao. 2023. Divide-and-conquer model based on wavelet domain for multi-focus image fusion. *Signal Processing: Image Communication* (2023).

[30] Yuchen Xian, Yunqiu Xu, Yang He, and Yi Yang. 2026. From 2D Grids to 1D Tokens: Reforming Shared Representations for Multimodal Image Fusion. In *ICML*.

[31] Housheng Xie, Meng Sang, Yukuan Zhang, Yang Yang, Shan Zhao, and Jianbo Zhong. 2024. RCVS: A Unified Registration and Fusion Framework for Video Streams. *IEEE TMM* (2024).

[32] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. 2020. U2Fusion: A Unified Unsupervised Image Fusion Network. *IEEE TPAMI* (2020).

[33] Xunpeng Yi, Linfeng Tang, Hao Zhang, Han Xu, and Jiayi Ma. 2024. Diff-IF: Multi-modality image fusion via diffusion model with fusion knowledge prior. *Information Fusion* (2024).

[34] Guanghui Yue, Wentao Li, Cheng Zhao, Zhiliang Wu, Tianwei Zhou, Qiuping Jiang, and Runmin Cong. 2025. Text-Guided Semantic Alignment Network With Spatial-Frequency Interaction for Infrared-Visible Image Fusion Under Extreme Illumination. *IEEE TIP* (2025).

[35] Cheng Zhao, Tianyun Song, Zhiliang Wu, Tianfu Wang, Moncef Gabbouj, Guanghui Yue, Baiying Lei, and Wei Zhou. 2026. Uncertainty-Guided Spatiotemporal Consistency Fusion Network for Infrared-Visible Video Fusion under Extremely Low-Light Conditions. *IEEE TIP* (2026).

[36] Ganlong Zhao, Guanbin Li, Yipeng Qin, and Yizhou Yu. 2023. Improved distribution matching for dataset condensation. In *CVPR*.

[37] Zixiang Zhao, Haowen Bai, Bingxin Ke, Yukun Cui, Lilun Deng, Yulun Zhang, Kai Zhang, and Konrad Schindler. 2025. A Unified Solution to Video Fusion: From Multi-Frame Learning to Benchmarking. In *NeurIPS*.

[38] Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jiangshe Zhang, Shuang Xu, Yulun Zhang, Kai Zhang, Deyu Meng, Radu Timofte, and Luc Van Gool. 2023. DDFM: Denoising Diffusion Model for Multi-Modality Image Fusion. In *ICCV*.

[39] Zixiang Zhao, Yukun Cui, Lilun Deng, Haowen Bai, Haotong Qin, Tao Feng, and Konrad Schindler. 2026. MambaVF: State Space Model for Efficient Video Fusion. *arXiv preprint arXiv:2602.06017* (2026).

[40] Kun Zhou, Wenbo Li, Liying Lu, Xiaoguang Han, and Jiangbo Lu. 2022. Revisiting Temporal Alignment for Video Restoration. In *CVPR*.

[41] Yabin Zhu, Qianwu Wang, Chenglong Li, Jin Tang, Chengjie Gu, and Zhixiang Huang. 2025. Visible-Thermal Multiple Object Tracking: Large-scale Video Dataset and Progressive Fusion Approach. *PR* (2025).