Fusing Thermal and Event Data for Visible Spectrum Image Reconstruction

Simone Melcarne[1] and Jean-Luc Dugelay[1]
[1]*Eurecom Research Center, Digital Security Department, Biot, France*

Keywords: Computational Imaging, Cross-spectral Vision, Multimodal Fusion, Thermal Infrared, Event-based Cameras.

Abstract: Reconstructing visible spectrum images from unconventional sensors is a timely and relevant problem in computer vision. In settings where standard cameras fail or are not allowed, thermal and event-based cameras can offer complementary advantages—robustness to darkness, fog, motion, and high dynamic range conditions—while also being privacy-preserving and energy efficient. However, their raw data is hard to read, and most computer vision models are designed and pretrained on standard visible inputs, making direct integration of unconventional data challenging. In this work, we ask whether it is possible, given a paired system that simultaneously records thermal and event data, to recover the kind of information people associate with the visible spectrum. We propose a simple dual-encoder, gated-fusion network that synthesizes visible-like images from thermal frames and event streams. The thermal branch captures structure and coarse appearance; the event branch models spatio-temporal changes and adds more detailed information. Their outputs are combined together and finally decoded into a colored image. We train and test the proposed solution on a paired thermal–visible–event dataset. Results show that this approach can recover plausible visible images producing better results than single-modality baselines, both quantitatively and qualitatively.

1 INTRODUCTION

Thermal infrared (TIR) and event-based cameras are widely used across many computer vision applications, where standard RGB imaging may not be reliable (*e.g.,* low light, fog, fast motion). In contrast, TIR sensors remain effective at night or through visual obstructions [Berg et al., 2015], while event-based cameras capture fine edge dynamics in scenes with rapid movement or extreme lighting [Gallego et al., 2022]. These properties make them appealing for robustness and efficiency. Furthermore, in line with the emerging concept of *diminished sensing* [CNIL, 2024] which promotes the use of sensors that deliberately reduce the type and amount of data collected, thermal and event cameras are privacy-aware by nature. However, both sensors present significant limitations when used in isolation. They typically suffer from low spatial resolution and lack the rich texture and color information available in RGB images. On top of that, the majority of computer vision models, particularly those for detection and recognition, are designed and pretrained on RGB inputs, making it difficult to directly integrate thermal or event data into standard pipelines. For these reasons, recent research has explored how to reconstruct visible content from each one independently. In the thermal-to-visible translation task, deep learning methods have evolved from simple gray-scale colorization to CNN-based models and more advanced Generative Adversarial Networks (GANs) [Goodfellow et al., 2014]. Translating events into visible images is an even more challenging conversion task due to the data's asynchronous and non-frame-based nature. While early approaches used classical pipelines with hand-crafted motion constraints, the advent of deep learning introduced data-driven methods that reconstruct video frames directly from event streams. We are not aware of prior work that has attempted to combine both thermal and event input for RGB synthesis, therefore we introduce a multimodal framework that fuses the two modalities to reconstruct visible color images. Our model uses a dual branch encoder in order to make thermal frames providing semantic layout and object presence, while events offering sharp edge and motion cues. The goal of this work is to show that fusing thermal infrared and event streams within a single model yields higher quality reconstructions than either modality alone. This design enables the idea of a new type of visual acquisition system that potentially avoids conventional RGB sensors and instead reconstructs images from reduced input. We evaluated quantitatively and qualitatively the proposed solution on a tri-modal dataset.

2 BACKGROUND

Thermal Vision The TIR portion of the electromagnetic spectrum represents the radiation emitted by objects. This emitted energy is directly linked to the temperature, with warmer objects releasing more infrared radiation. The TIR band is often categorized into mid-wave (MWIR, 3-5 μm) and long-wave infrared (LWIR, 8-12 μm), with LWIR being particularly useful for night vision, since everyday objects at typical temperatures emit most of their thermal radiation in this range. Unlike conventional cameras that capture reflected light, the data recorded by these cameras consists of thermal profiles, where each pixel corresponds to the intensity of infrared radiation, effectively creating a heat map of the observed area.

Event Vision Event based cameras are a new kind of vision sensor inspired by how biological eyes work [Gallego et al., 2022, Adra et al., 2025]. Unlike traditional cameras that capture full images at fixed time intervals, event cameras work asynchronously, *i.e.,* each pixel reacts on its own and records an event only when it detects a significant change in brightness. Specifically, an event e is generated at a certain pixel $\mathbf{u} = (u, v)$ whenever the logarithmic change of the irradiance R calculated between the current time t and the time $t - \delta t$ of occurrence of the last event, exceeds a predefined positive contrast threshold Θ:

$$e(\mathbf{u}, t, \sigma) := \log(R(\mathbf{u}, t)) - \log(R(\mathbf{u}, t - \delta t)) \geq \Theta \quad (1)$$

Each event encodes the pixel's identity $\mathbf{u} = (u, v)$ that indicate the location of the change, a timestamp t, capturing the precise time the event occurred, and the polarity $\sigma \in \{-1, +1\}$, specifying whether the brightness increased or decreased.

2.1 Related Work

Thermal to Visible Reconstruction Unlike standard grayscale colorization methods, the process of converting from TIR domain to RGB must involve the recover of both luminance and chrominance, since thermal measurements do not align with visible appearance [Berg et al., 2018]. Early work often treated the thermal frame as a pseudo-luminance and predicted only chroma, which limited the realism [Cao et al., 2017, Deshpande et al., 2017, Guadarrama et al., 2017]. Research later moved from CNNs to generative and attention-based models, often using UNet-style encoder-decoder architectures to map a single TIR frame to a color space [Berg et al., 2018]. The adoption of GANs, particularly Pix2Pix [Isola et al., 2017] with paired supervision, significantly improved texture and color, with variants refining the generator to sharpen edges and recover small structures [Liao et al., 2023, Mirza and Osindero, 2014]. More recent methods [Jiang et al., 2025, Zhan et al., 2025, He et al., 2023] deliver high quality results, but they rely on deep and complex neural networks and require extensive training on large image datasets to learn the mapping between grayscale and color images.

Event to Visible Reconstruction Due to the sparse and asynchronous nature of event data, which is hard to visualize and incompatible with standard neural architectures, a line of research has focused on reconstructing luminance-like images from events. Early methods relied on gradient accumulation and Poisson integration under static scene assumptions [Cook et al., 2011, Kim et al., 2008]. Subsequent methods handled generic motion by jointly estimating intensity and optical flow via variational optimization [Bardow et al., 2016]. More recent works adopt deep learning, with E2VID [Rebecq et al., 2019a, Rebecq et al., 2019b] being a reference model that reconstructs grayscale video from raw events. From here, several variants have been proposed to improve efficiency [Scheerlinck et al., 2018, Stoffregen et al., 2020]. While most methods focus on grayscale reconstruction, recent trends have explored the possibility to combine events with RGB baseline frames to perform image enhancement in high dynamic range scenarios [Shaw et al., 2022, Han et al., 2023, Yang et al., 2023, Cui et al., 2024].

3 METHODOLOGY

We base our work on the hypothesis of a dual-camera recording system where both devices operate at the same time. On the one hand, the thermal camera captures an image $T \in \mathbb{R}^{1 \times H \times W}$ where the single channel corresponds to grayscale intensity while H and W denote the height and width of the image, respectively. On the other hand, the event-based camera generates a continuous stream of events $\mathcal{E}_\mathcal{K} = \{e_k\}_{k=1}^{K}$, with K representing the total number of recorded events during the exposure time of the thermal image T. To process asynchronous event streams using CNNs, we convert the data into a fixed-size voxel grid representation that encodes temporal information in a B-channels 3D volume [Zhu et al., 2018] [Han et al., 2023]. Given a sequence of K events $e_k = (\mathbf{u}_k, t_k, \sigma_k)$, we discretize the temporal window $\Delta t = [t_0, t_{K-1}]$ into B bins. Equation 2 evenly assigns the entire stream $\mathcal{E}_\mathcal{K}$ into B temporal bins t:

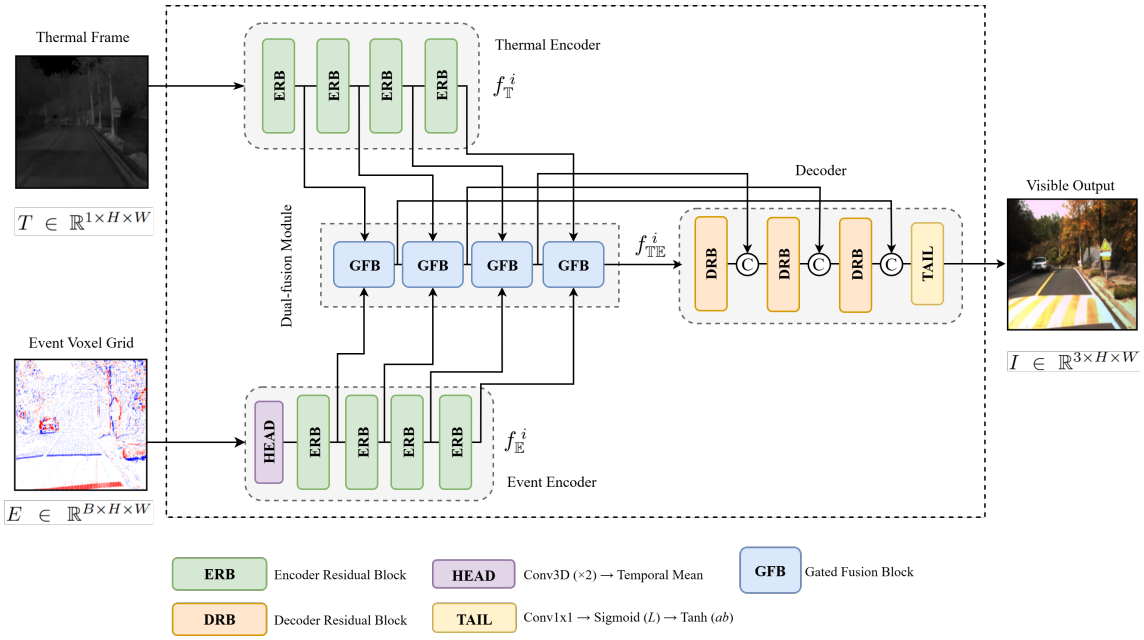

Figure 1: Overview of the proposed framework. The network receives both the thermal image and the event voxel grid as inputs. A dual encoder setup extracts meaningful features from both modalities. The final visible like output is fused by a scalar residual gated fusion block and reconstructed by the shared decoder.

$$E_t = \sum_{k=0}^{K-1} \sigma_k \max(0,\, 1 - |\, t - \hat{t}_k \,|) \qquad (2)$$
$$\text{with } t = 0,\ldots,B-1, \quad \hat{t}_k = \frac{t_k - t_0}{\Delta t}(B-1).$$

Considering a thermal frame $T \in \mathbb{R}^{1\times H\times W}$ and a voxel grid sequence of event data $E \in \mathbb{R}^{B\times H\times W}$, our objective is to reconstruct a 3-channels color image $I \in \mathbb{R}^{3\times H\times W}$ using a multimodal learning framework.

3.1 Network Architecture

The proposed architecture consists of two parallel encoders designed to extract features from thermal frames and event voxel grids. Both branches produce pyramids of modality features using residual blocks, which are fused at each level and passed to a shared decoder that reconstructs the final visible output. The pipeline is summarized in Figure 1.

3.1.1 Thermal Encoder

The thermal encoder receives a single-channel thermal image $T \in \mathbb{R}^{1\times H\times W}$ and processes it through a 4-stage residual block pyramid (ERB in Figure 1) [Mei et al., 2020, He et al., 2015, Lim et al., 2017]. Each block consists of two convolutional layers with 3×3 kernels, followed by group normalization [Wu and He, 2018] and ReLU activations with dropout. After the first three blocks, a max pooling layer halves the spatial resolution, resulting in feature maps of size 128×128, 64×64, and 32×32, respectively. The number of channels increases progressively across the encoder: the first block outputs 32 channels, the second 64, then 128, and finally 256 in the deepest layer. The encoder outputs the four tensors $f_{\mathbb{T}}^{1},..,f_{\mathbb{T}}^{4}$. This fine-to-coarse hierarchy captures global context and sharp boundaries, which is particularly helpful to disambiguate color from structure when fusing thermal cues with sparse event features.

3.1.2 Event Encoder

The event encoder operates in parallel and processes a B-channel voxel grid $E \in \mathbb{R}^{B\times H\times W}$, where each channel is a temporal bin of accumulated events. To exploit the spatio-temporal structure while suppressing bin-level noise, we first employ a lightweight 3D convolutional block [Tran et al., 2015], where two 3×3×3 Conv3D layers, followed by group normalization and ReLU, mix time and space and produce 16 feature maps, which are then collapsed over time by averaging to finally obtain a 2D representation (HEAD in Figure 1). This is consistent with event-vision practice of aggregating events into grid representations for processing with standard 2D CNNs [Gehrig et al., 2019, Rebecq et al., 2019a]. From here the encoder mirrors the thermal pyramidal structure

Figure 2: Dataset examples from the training set. From left to right the thermal infrared (TIR) frame, the temporal bin-wise sum projection image from the event voxel grid, and the RGB ground truth.

of the encoder producing $f_{\mathbb{E}}^1, .., f_{\mathbb{E}}^4$ with 64, 128, and 256 channels at 128×128, 64×64, and 32×32 spatial resolutions. In this way, both thermal and event encoders produce features at each level with matched scale and channel counts, allowing for a simple and direct fusion stage.

3.1.3 Dual-Fusion Module

At each resolution level, we fuse thermal and event features with a feature-level residual gated addition (GFB in Figure 1):

$$f_{\mathbb{TE}}^{(i)} = f_{\mathbb{T}}^{(i)} + \alpha_i \cdot f_{\mathbb{E}}^{(i)}, \tag{3}$$

where α_i is the learned gate at scale i, and $f_{\mathbb{T}}^{(i)}$, $f_{\mathbb{E}}^{(i)}$ are the corresponding features from the thermal and event encoders. Compared to data- or output-level fusion, this feature-level integration is often more effective because it enables deeper inter-modal interaction with respect to superficial concatenation [Li and Tang, 2024], allowing the model to use events only when helpful and always retaining a thermal baseline [Hazirbas et al., 2016].

3.1.4 Decoder

The shared decoder takes the four fused features $f_{\mathbb{TE}}^1, .., f_{\mathbb{TE}}^4$ and reconstructs in three U-Net–style [Ronneberger et al., 2015] stages. Starting from $f_{\mathbb{TE}}^4$, each stage upsamples with a 2×2 transposed convolution reducing progressively from 256 to 32 channels, concatenates the result with the fused skip at the same scale and finally refines with the residual block (DRB in Figure 1) in order to reduce channels back to 128, 64, and 32, respectively. These skip connections help restore fine-grained spatial structure and motion cues that were lost during the downsampling phase. As last step (TAIL in Figure 1), a 1×1 convolution outputs three channels in *Lab* space where L is passed through a sigmoid ($L = \sigma(\cdot) \in [0,1]$) and ab through a tanh function ($ab = \tanh(\cdot) \in [-1,1]$). Since the objective is to reconstruct colored images, we formulated the problem in the *Lab* color space as commonly chosen when dealing with tasks that involve minimizing color perceptual differences [Zhang et al., 2016, Berg et al., 2018, Iizuka et al., 2016].

3.2 Objective Function

To train the proposed framework in an end-to-end fashion, we made use of a combination of three different loss functions:

$$\mathcal{L}_{\text{total}} = \alpha\,\mathcal{L}_{s-\ell_1} + \beta\,\mathcal{L}_{\text{MS-SSIM}} + \gamma\,\mathcal{L}_{ab-\ell_1}, \tag{4}$$

with α, β, γ empirically set to $0.6, 0.4, 0.8$ respectively. The $\mathcal{L}_{\text{s-}\ell_1}$ loss [1] is computed on RGB domain after color space transfromation from *Lab* and minimizes per-pixel photometric error for each channel, remaining robust to outliers. Given the ground truth $I \in \mathbb{R}^{3\times H\times W}$, and the reconstruction $\hat{I} \in \mathbb{R}^{3\times H\times W}$, the loss is defined as:

$$\mathcal{L}_{\text{s-}\ell_1} = \frac{1}{N}\sum_{p\in\Omega} \rho_\delta\big(I_p - \hat{I}_p\big),$$
$$\rho_\delta(d) = \begin{cases} \dfrac{d^2}{2\delta}, & |d| < \delta, \\ |d| - \dfrac{\delta}{2}, & \text{otherwise.} \end{cases} \tag{5}$$

where $p = (x,y)$ is a pixel, $\Omega = \{1,\dots,H\} \times \{1,\dots,W\}$ the set of pixels, $N = |\Omega|$ its cardinality, and δ a threshold set to 0.01. The $\mathcal{L}_{\text{MS-SSIM}}$ [2] loss is computed on luminance and it is designed to preserve perceived structure and contrast across multiple scales (*i.e.,* MS-SSIM). Given $L \in \mathbb{R}^{1\times H\times W}$ the ground truth luminance map and $\hat{L} \in \mathbb{R}^{1\times H\times W}$ the reconstructed one:

$$\mathcal{L}_{\text{MS-SSIM}} = \frac{1}{N}\sum_{p\in\Omega} 1 - \frac{\text{MS-SSIM}(L_p, \hat{L}_p))}{2}, \tag{6}$$

with the local SSIM generally defined as

$$\text{SSIM}(X,Y) = \frac{(2\mu_X\mu_Y + C_1)\,(2\sigma_{XY} + C_2)}{(\mu_X^2 + \mu_Y^2 + C_1)\,(\sigma_X^2 + \sigma_Y^2 + C_2)}, \tag{7}$$

where μ, σ^2, and σ_{XY} are respectively local mean, variance, and covariance; $C_1, C_2 > 0$ are pre-defined

[1] Pytorch implementation: *Smooth* L1

[2] Kornia implementation: SSIM

Table 1: Quantitative comparison with mean and standard deviation on KAIST-MS validation and test sets. Best in **bold**.

Method	PSNR↑		SSIM↑		LPIPS↓	
	val	*test*	*val*	*test*	*val*	*test*
TIR2Lab	17.02 ±2.83	14.68 ±3.05	0.617 ±0.12	0.544 ±0.13	0.448 ±0.12	0.501 ±0.12
E2VID	10.62 ±1.67	11.17 ±1.20	0.450 ±0.09	0.434 ±0.10	0.398 ±0.11	0.382 ±0.08
Ours (*thermal*)	16.00 ±2.71	16.47 ±2.46	0.534 ±0.12	0.553 ±0.13	0.460 ±0.08	0.436 ±0.10
Ours (*event*)	17.75 ±3.22	17.63 ±3.47	0.687 ±0.14	0.711 ±0.10	0.316 ±0.11	0.277 ±0.08
Ours (*full*)	**19.17 ±2.39**	**18.88 ±2.20**	**0.732 ±0.07**	**0.752 ±0.07**	**0.266 ±0.06**	**0.241 ±0.06**

Table 2: Ablation of loss terms and fusion strategies. Best in **bold**.

Method	PSNR↑		SSIM↑		LPIPS↓	
	val	*test*	*val*	*test*	*val*	*test*
$\mathcal{L}_{s-\ell_1}$	18.48 ±2.54	18.71 ±2.71	0.708 ±0.08	0.721 ±0.09	0.287 ±0.06	0.259 ±0.06
$+\mathcal{L}_{\text{MS-SSIM}}$	18.89 ±2.35	18.47 ±2.29	0.723 ±0.08	0.736 ±0.07	0.273 ±0.06	0.257 ±0.06
$+\mathcal{L}_{ab-\ell_1}$ (*full*)	**19.17 ±2.39**	**18.88 ±2.20**	**0.732 ±0.07**	**0.752 ±0.07**	**0.266 ±0.06**	**0.241 ±0.06**
Add	17.47 ±2.74	18.18 ±2.37	0.652 ±0.08	0.691 ±0.09	0.328 ±0.07	0.283 ±0.07
Concat	15.41 ±2.58	16.37 ±2.71	0.514 ±0.12	0.525 ±0.13	0.480 ±0.08	0.461 ±0.10
Gated residual	**19.17 ±2.39**	**18.88 ±2.20**	**0.732 ±0.07**	**0.752 ±0.07**	**0.266 ±0.06**	**0.241 ±0.06**

constants. The $\mathcal{L}_{ab-\ell_1}$ color loss enforces the network to hallucinate colors penalizing neutral washed-out colors. Let $(a_p, b_p), (\hat{a}_p, \hat{b}_p) \in \mathbb{R}^{2\times H\times W}$ be the ground-truth and predicted *Lab* chroma channels, the loss is defined as:

$$\mathcal{L}_{ab-\ell_1} = \frac{1}{N}\sum_{p\in\Omega}\Big(|\hat{a}_p - a_p| + |\hat{b}_p - b_p|\Big), \quad (8)$$

with p, Ω and N as previously defined.

4 EXPERIMENTAL SETUP

4.1 Dataset

To develop and evaluate our proposed framework, we require a tri-modal dataset consisting of synchronized paired triplets: thermal, event, and RGB data, with the RGB modality serving as the reference target. We used the day-set of KAIST-MS dataset [Hwang et al., 2015] which features RGB-TIR images captured from a moving vehicle on daylight. Events are synthetically generated from RGB using an event simulator, which output stream of events in (t, x, y, σ) format. In our experiments we choose the V2E [Hu et al., 2021] simulator. We converted then these streams into spatio-temporal voxel grids as described in Equation 2. Each voxel grid was generated by accumulating events within the timestamp range of its corresponding RGB frame, allowing to obtain a 1-1-1 match across the three modalities. We set $B = 5$, a standard choice in the literature [Shaw et al., 2022, Han et al., 2023, Weng et al., 2024]. The dataset is composed of sets which contain multiple videos, so we partitioned the data at the video level within each set. Every triplet from a sequence is assigned to exactly one split (train, validation or test). We made this choice because a random frame-level split would leak near duplicates across the splits. In total there are 28324 triplets, of which 17579 are used for training, 9058 for validation and 1687 for test. Figure 2 contains samples randomly extracted from the training set.

4.2 Implementation Details

During training, thermal and RGB images are resized to 1×256×256 and 3×256×256, respectively, and the corresponding input event voxel grids to 5×256×256. To prevent overfitting we applied minor transformations for data augmentation, consisting of a random crop and flip. We implemented our model in PyTorch, performing the training on a `NVIDIA GeForce RTX 3090 GPU` for 100 epochs using a batch size of 16, a learning rate of $3e^{-4}$ with `Adam` optimizer and `ReduceLROnPlateau` scheduler, with a patience of 5 epochs and a reduction factor of 0.5, along with an early stopping criterion.

5 RESULTS

Evaluation Metrics In this section we present the quantitative evaluation. We made use of three widely-recognized metrics for image reconstruction tasks: Peak Signal-to-Noise Ratio (PSNR), where higher values suggest better performance in terms of absolute fidelity; Structural Similarity Index (SSIM), which quantifies degradation as a perceived change in structural information, brightness, and contrast (higher SSIM denotes higher quality); and Learned Perceptual Image Patch Similarity (LPIPS), a perceptual metric that uses a deep neural network to measure

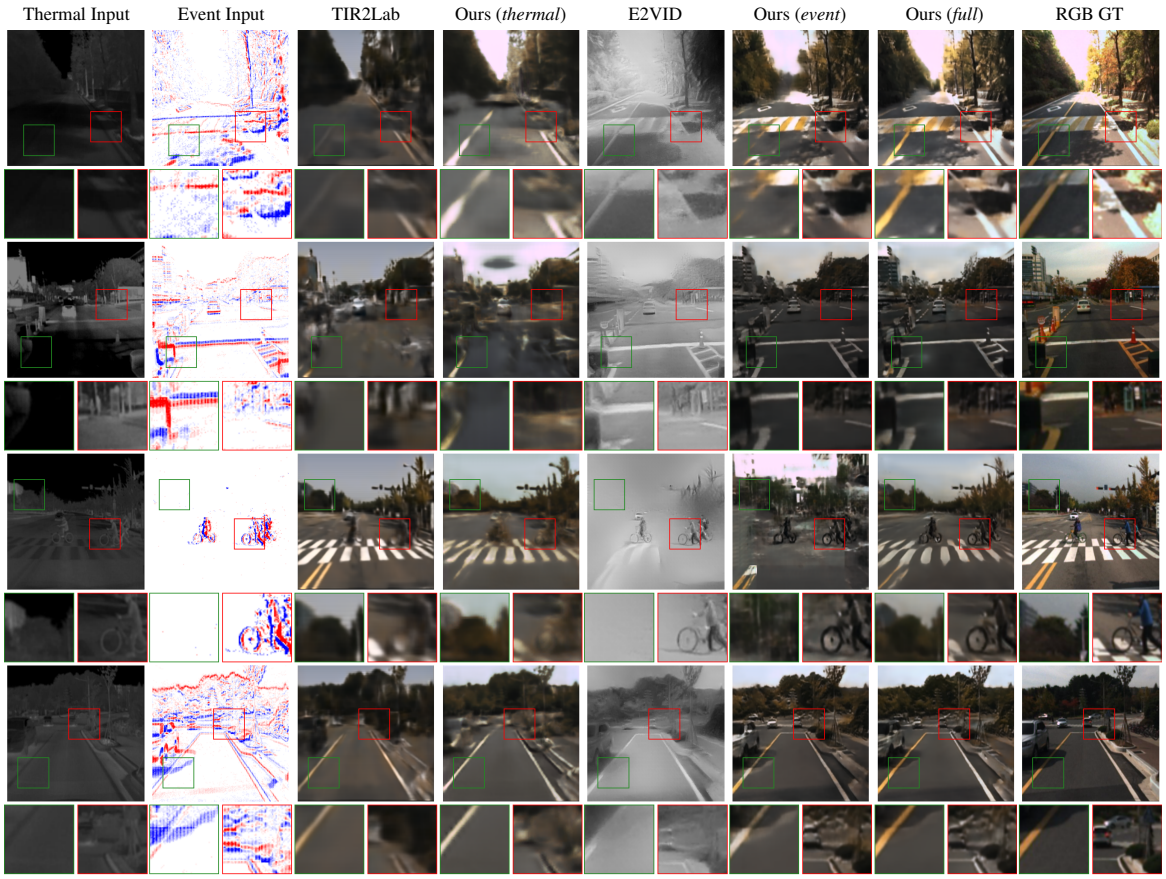

Figure 3: Visual comparison example. We present a comparison between images produced with each specific variant of the proposed solution and with two different single-modality representative works.

differences in the feature space (lower LPIPS is better) [Zhang et al., 2018].

Ablation and Comparison We evaluate the proposed model conducting an ablation study to assess the contribution of each modality, considering three variants: *full*, *thermal*, and *event*. We also compute the scores for two other single-modality seminal frameworks (TIR2Lab) [Berg et al., 2018] and (E2VID) [Rebecq et al., 2019a], respectively, from thermal-to-visible and event-to-visible. We want to highlight that our objective is not to surpass specialized translation systems that target perceptual realism but to show that the joint use of thermal and event improves the recovery of visible images. This explains why we compare against two foundational baselines rather than against more sophisticated GAN-based frameworks that already deliver strong single-modality translations. Table 1 reports the results on the validation and test sets of the KAIST-MS data set [Hwang et al., 2015]. Note that E2VID produces gray-scale images from a given stream of events; therefore, we computed the metrics using gray-scale ground truth in this case. In Figure 3 we show a qualitative comparison for all the methods. We can notice how thermal-only baselines often deliver relatively coarse results, with loss of detail particularly around edges, making the addition of information from events useful (*e.g., first row*). In contrast, in situations where events are highly sparse and not sufficiently informative, the thermal modality provides a robust backbone that mitigates the failure modes of event-only approaches, leading to more acceptable reconstructions (*e.g., third row*). Figure 4 shows features representations from the thermal and event encoders, together with the fused map. Thermal features provide coarse structure and smooth regions, while event features emphasize edges and motion boundaries, suggesting that the fusion process effectively acts as a Fourier spectrum synthesis, where the thermal stream provides the low-frequency content and the events recover the high-frequency part. The proposed gated residual fusion balances this complementary information, resulting in a fused map that main-

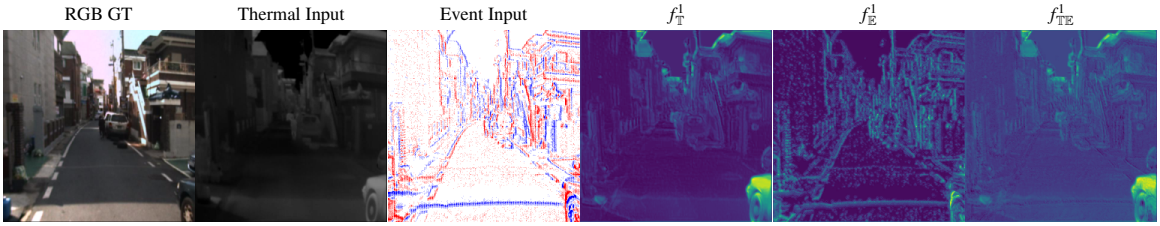

Figure 4: Feature maps visualization. Thermal, event, and fused feature maps at the first stage (finest resolution) are shown.

tains global consistency and enhanced contours. Finally, in Table 2 we first report the effects of each single term in the total loss formulation, showing that adding $\mathcal{L}_{\text{MS-SSIM}}$ and then $\mathcal{L}_{ab-\ell_1}$ improves the overall metric scores. Similarly, in the bottom part of the table we present three different fusion strategies. *Add* denotes plain feature summation ($f_{\mathbb{T}} + f_{\mathbb{E}}$), while *Concat* stacks features channel-wise ($f_{\mathbb{T}} || f_{\mathbb{E}}$). *Gated residual*, as defined in Equation 3, outperforms *Add* and *Concat* since the scale-wise gate selectively injects event cues, avoiding destructive fusion results.

6 LIMITATIONS

While the proposed approach shows that events and thermal complement each other to improve reconstruction, we acknowledge that failure cases remain. In general, colorization from data that do not encode chromatic cues is an intrinsically under-determined problem, leaving the possibility to produce inconsistent colors. The main limitation lies in the use of simulated events. While this ensures tight alignment with the ground truth RGB and allows controlled experiments, real event data might introduce different noise profiles and dynamics. Addressing this gap with real multimodal recordings would strengthen the conducted analysis.

7 CONCLUSIONS

In this work, we tackled the problem of generating visible color images from two unconventional sensors, *i.e.,* thermal and event-based data, without relying on a standard camera. We proposed a simple dual-encoder with gated fusion network, capturing useful and complementary cues from each data modality. Preliminary results showed that, when combined together, the quality of the reconstructed outputs improves, especially around edges and detailed regions. As we used simulated event streams, it is important to assess the performance on a real-world dataset. In this sense, one solid contribution would be to build a tri-modal real-world dataset with synchronized data of the three modalities. Finally, the current model operates frame-by-frame without enforcing temporal coherence; future work can incorporate temporal consistency constraints.

ACKNOWLEDGEMENTS

This work was partly supported by the European Union's Horizon Europe research and innovation program under Grant Agreement No 101094831 for the Converge-Telecommunications and Computer Vision Convergence Tools for Research Infrastructures project. The authors state that AI tools have been used for language assistance. No AI tool was used to generate or modify research data, figures, or results.

REFERENCES

Adra, M., Melcarne, S., Mirabet-Herranz, N., and Dugelay, J.-L. (2025). Event-based solutions for human-centered applications: a comprehensive review. *Frontiers in Signal Processing*, Volume 5 - 2025.

Bardow, P., Davison, A. J., and Leutenegger, S. (2016). Simultaneous optical flow and intensity estimation from an event camera. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 884–892.

Berg, A., Ahlberg, J., and Felsberg, M. (2015). A thermal object tracking benchmark. In *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6.

Berg, A., Ahlberg, J., and Felsberg, M. (2018). Generating visible spectrum images from thermal infrared. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1224–1233.

Cao, Y., Zhou, Z., Zhang, W., and Yu, Y. (2017). Unsupervised diverse colorization via generative adversarial networks. In Ceci, M., Hollmén, J., Todorovski, L., Vens, C., and Džeroski, S., editors, *Machine Learning and Knowledge Discovery in Databases*, volume 10534 of *Lecture Notes in Computer Science*, pages 151–166, Cham. Springer.

CNIL (2024). De la vidéo augmentée à des capteurs diminués. Accessed: 2025-06-27.

Cook, M., Gugelmann, L., Jug, F., Krautz, C., and Steger, A. (2011). Interacting maps for fast visual interpretation. In *The 2011 International Joint Conference on Neural Networks*, pages 770–776. IEEE.

Cui, M., Wang, Z., Wang, D., Zhao, B., and Li, X. (2024). Color event enhanced single-exposure hdr imaging. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(2):1399–1407.

Deshpande, A., Lu, J., Yeh, M.-C., Jin Chong, M., and Forsyth, D. (2017). Learning diverse image colorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gallego, G., Delbruck, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A. J., Conradt, J., Daniilidis, K., and Scaramuzza, D. (2022). Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180.

Gehrig, D., Loquercio, A., Derpanis, K. G., and Scaramuzza, D. (2019). End-to-end learning of representations for asynchronous event-based data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5633–5643.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial networks.

Guadarrama, S., Dahl, R., Bieber, D., Norouzi, M., Shlens, J., and Murphy, K. (2017). Pixcolor: Pixel recursive colorization. *CoRR*, abs/1705.07208.

Han, J., Yang, Y., Duan, P., Zhou, C., Ma, L., Xu, C., Huang, T., Sato, I., and Shi, B. (2023). Hybrid high dynamic range imaging fusing neuromorphic and conventional images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8553–8565.

Hazirbas, C., Ma, L., Domokos, C., and Cremers, D. (2016). Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Asian conference on computer vision*, pages 213–228. Springer.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

He, Y., Xin, J., Jiang, Q., Cheng, Z., Wang, P., and Zhou, W. (2023). Lkat-gan: A gan for thermal infrared image colorization based on large kernel and attentionunet-transformer. *IEEE Transactions on Consumer Electronics*, PP:1–1.

Hu, Y., Liu, S.-C., and Delbruck, T. (2021). v2e: From video frames to realistic DVS events. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. arXiv:2006.07722.

Hwang, S., Park, J., Kim, N., Choi, Y., and Kweon, I. S. (2015). Multispectral pedestrian detection: Benchmark dataset and baseline. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1037–1045.

Iizuka, S., Simo-Serra, E., and Ishikawa, H. (2016). Let there be color!: Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics (Proc. of SIGGRAPH)*, 35(4):110:1–110:11. Article 110.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.

Jiang, Q., Yao, G., Feng, M., Jin, X., Miao, S., Gao, Y., and Cheng, X. (2025). Mcu-gan: Colorization method for infrared images based on multi-convolution fusion and generative adversarial network. *Infrared Physics & Technology*, 145:105673.

Kim, H., Handa, A., Benosman, R., Ieng, S.-H., and Davison, A. J. (2008). Simultaneous mosaicing and tracking with an event camera. *J. Solid State Circ*, 43:566–576.

Li, S. and Tang, H. (2024). Multimodal alignment and fusion: A survey. *arXiv preprint arXiv:2411.17040*.

Liao, H., Jiang, Q., Jin, X., Liu, L., Liu, L., Lee, S.-J., and Zhou, W. (2023). Mugan: Thermal infrared image colorization using mixed-skipping unet and generative adversarial network. *IEEE Transactions on Intelligent Vehicles*, 8(4):2954–2969.

Lim, B., Son, S., Kim, H., Nah, S., and Lee, K. M. (2017). Enhanced deep residual networks for single image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Mei, Y., Fan, Y., Zhang, Y., Yu, J., Zhou, Y., Liu, D., Fu, Y., Huang, T. S., and Shi, H. (2020). Pyramid attention networks for image restoration. *arXiv preprint arXiv:2004.13824*.

Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Rebecq, H., Ranftl, R., Koltun, V., and Scaramuzza, D. (2019a). Events-to-video: Bringing modern computer vision to event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rebecq, H., Ranftl, R., Koltun, V., and Scaramuzza, D. (2019b). High speed and high dynamic range video with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Early Access.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation.

Scheerlinck, C., Barnes, N., and Mahony, R. (2018). Continuous-time intensity estimation using event cameras. In *Asian Conference on Computer Vision*, pages 308–324. Springer.

Shaw, R., Catley-Chandar, S., Leonardis, A., and Perez-Pellitero, E. (2022). Hdr reconstruction from bracketed exposures and events.

Stoffregen, T., Scheerlinck, C., Scaramuzza, D., Drummond, T., Barnes, N., Kleeman, L., and Mahony, R. (2020). Reducing the sim-to-real gap for event cameras. *arXiv preprint arXiv:2003.09078*.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497.

Weng, J., Li, B., and Huang, K. (2024). Event-based image enhancement under high dynamic range scenarios. In *Proceedings of the Asian Conference on Computer Vision*, pages 2456–2470.

Wu, Y. and He, K. (2018). Group normalization.

Yang, Y., Han, J., Liang, J., Sato, I., and Shi, B. (2023). Learning event guided high dynamic range video reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13924–13934.

Zhan, W., Chen, Y., Jiang, Y., Zhu, D., Xu, X., Guo, J., Hao, Z., Han, D., and Li, J. (2025). Reference-based infrared image colorization via feature enhancement and context refinement. *Knowledge-Based Systems*, 325:113982.

Zhang, R., Isola, P., and Efros, A. A. (2016). Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric.

Zhu, A. Z., Yuan, L., Chaney, K., and Daniilidis, K. (2018). Unsupervised event-based learning of optical flow, depth, and egomotion.