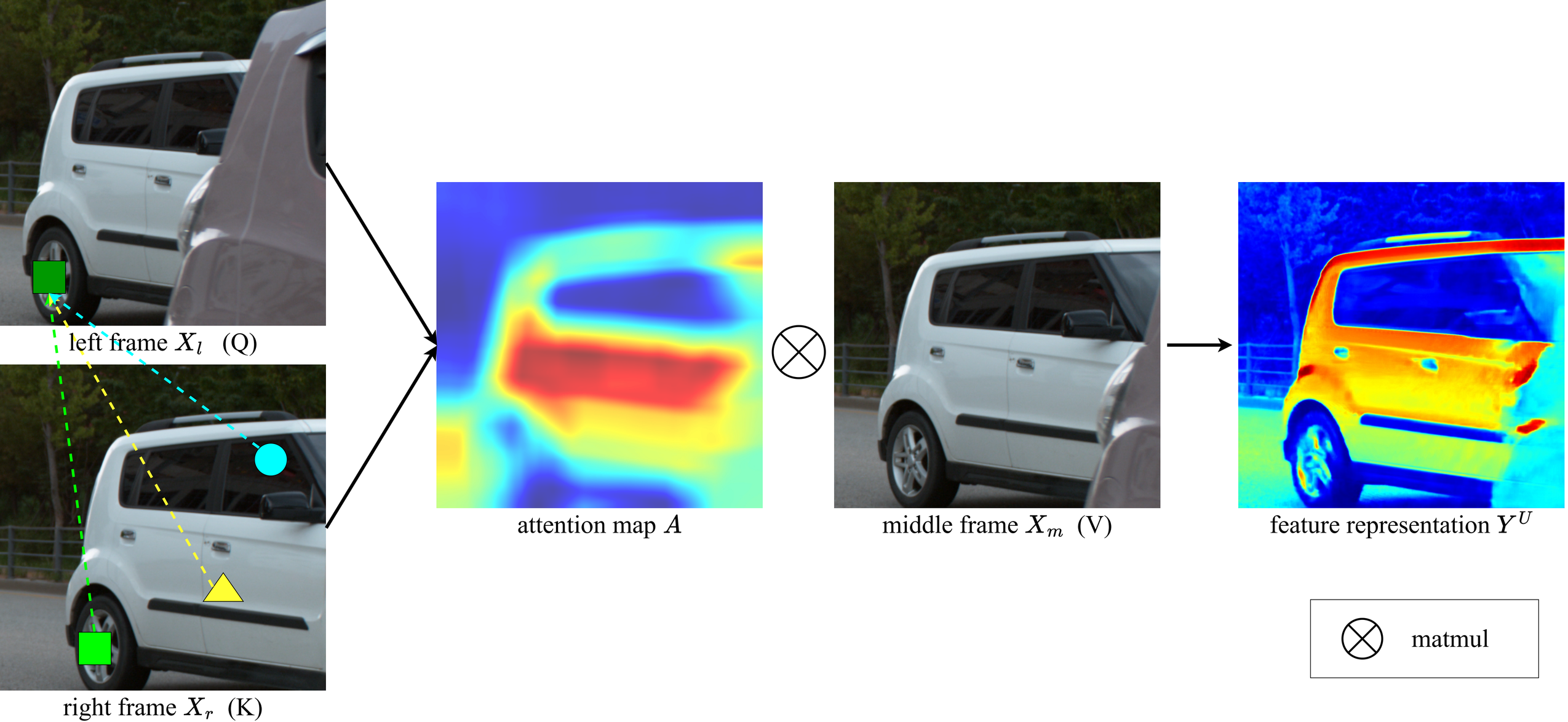
left frame X_l (Q)
right frame X_r (K)
attention map A
middle frame X_m (V)
feature representation Y^U
matmul