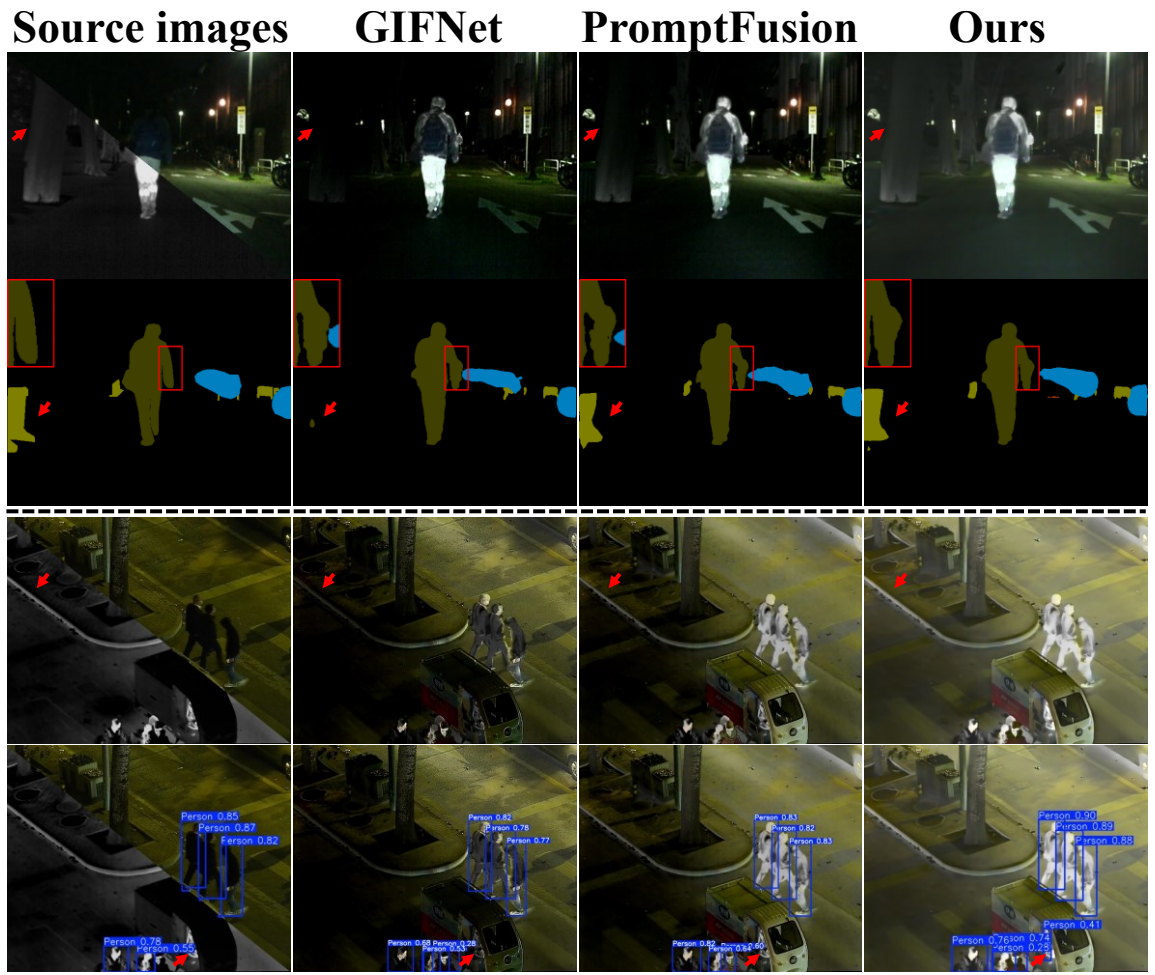
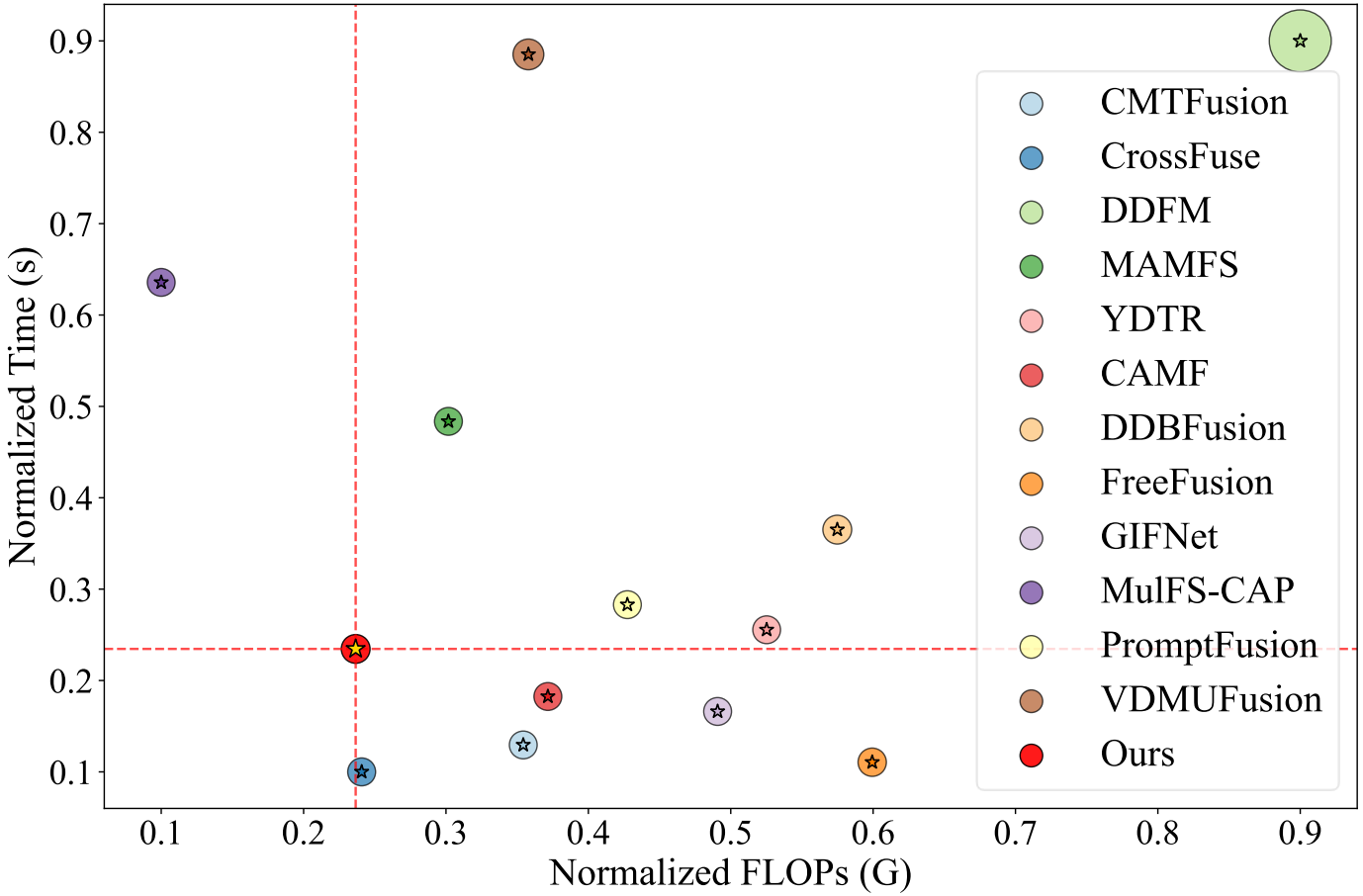
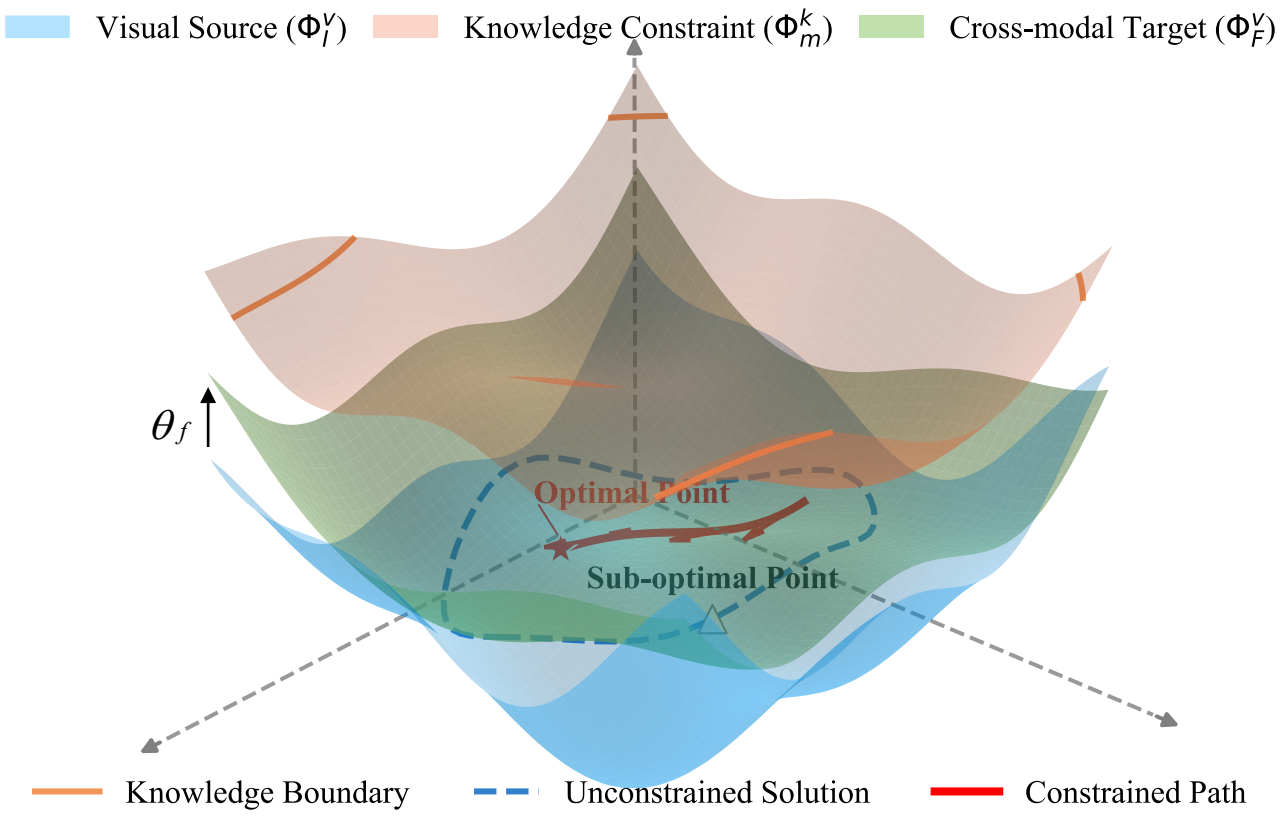
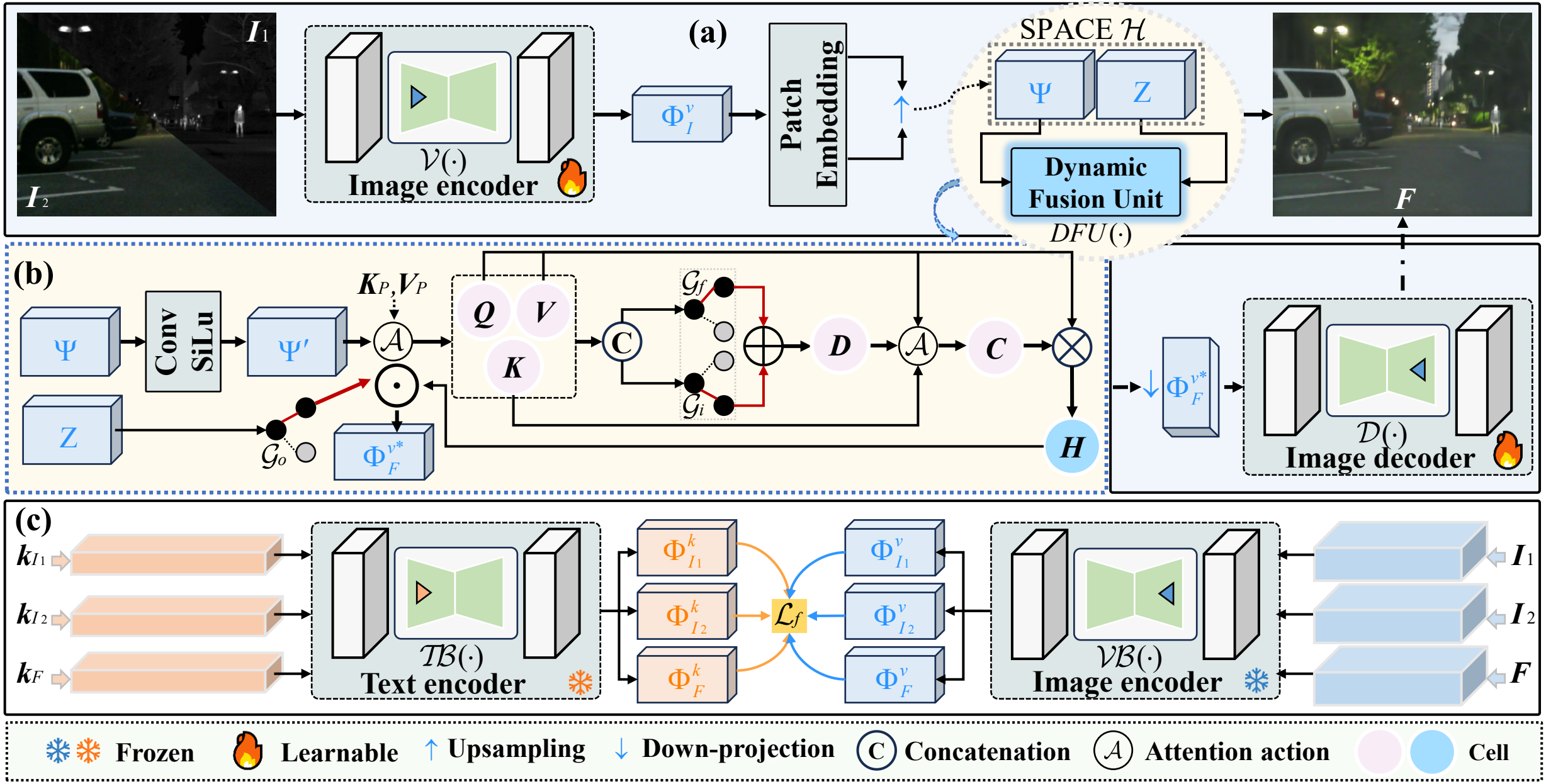
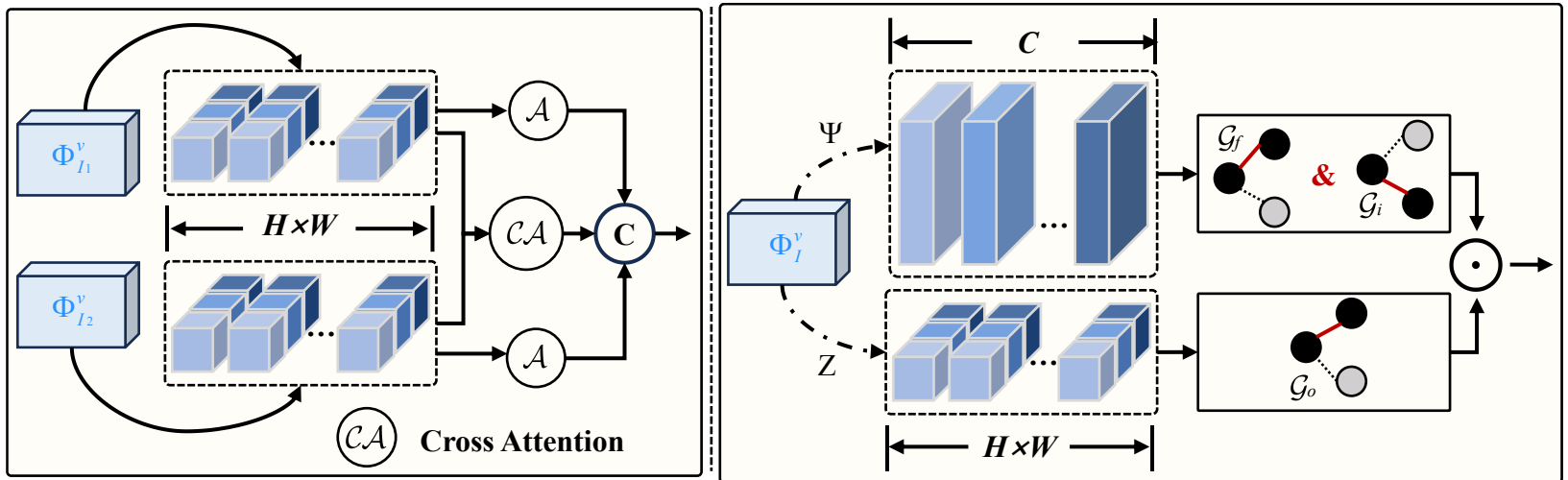

Source images
GIFNet
PromptFusion
Ours


Normalized Time (s)
Normalized FLOPs (G)
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
CMTFusion
CrossFuse
DDFM
MAMFS
YDTR
CAMF
DDBFusion
FreeFusion
GIFNet
MulFS-CAP
PromptFusion
VDMUFusion
Ours

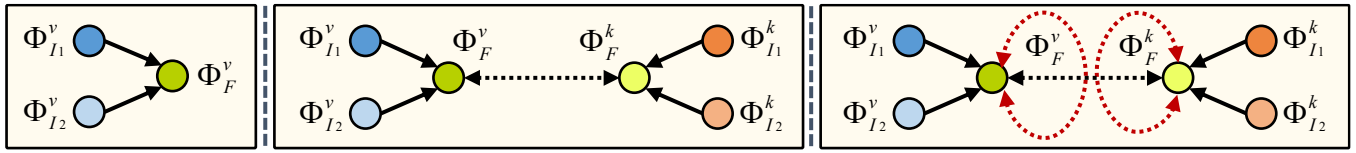

Visual Source (Φ_I^v)
Knowledge Constraint (Φ_m^k)
Cross-modal Target (Φ_F^v)
θ_f
Optimal Point
Sub-optimal Point
Knowledge Boundary
Unconstrained Solution
Constrained Path


(a)
I_1
I_2
$\mathcal{V}(\cdot)$
Image encoder
Φ_I^v
Patch Embedding
SPACE $\mathcal{H}$
Ψ
Z
Dynamic Fusion Unit
$DFU(\cdot)$
F
(b)
Ψ
Conv SiLu
Ψ'
Z
K_P, V_P
$\mathcal{A}$
Q
V
K
C
$\mathcal{G}_f$
$\mathcal{G}_i$
$\mathcal{G}_o$
D
$\mathcal{A}$
C
H
Φ_F^{v*}
$\mathcal{D}(\cdot)$
Image decoder
(c)
k_{I1}
k_{I2}
k_F
$\mathcal{TB}(\cdot)$
Text encoder
Φ_{I1}^k
Φ_{I2}^k
Φ_F^k
$\mathcal{L}_f$
Φ_{I1}^v
Φ_{I2}^v
Φ_F^v
$\mathcal{VB}(\cdot)$
Image encoder
I_1
I_2
F
Frozen
Learnable
Upsampling
Down-projection
Concatenation
Attention action
Cell

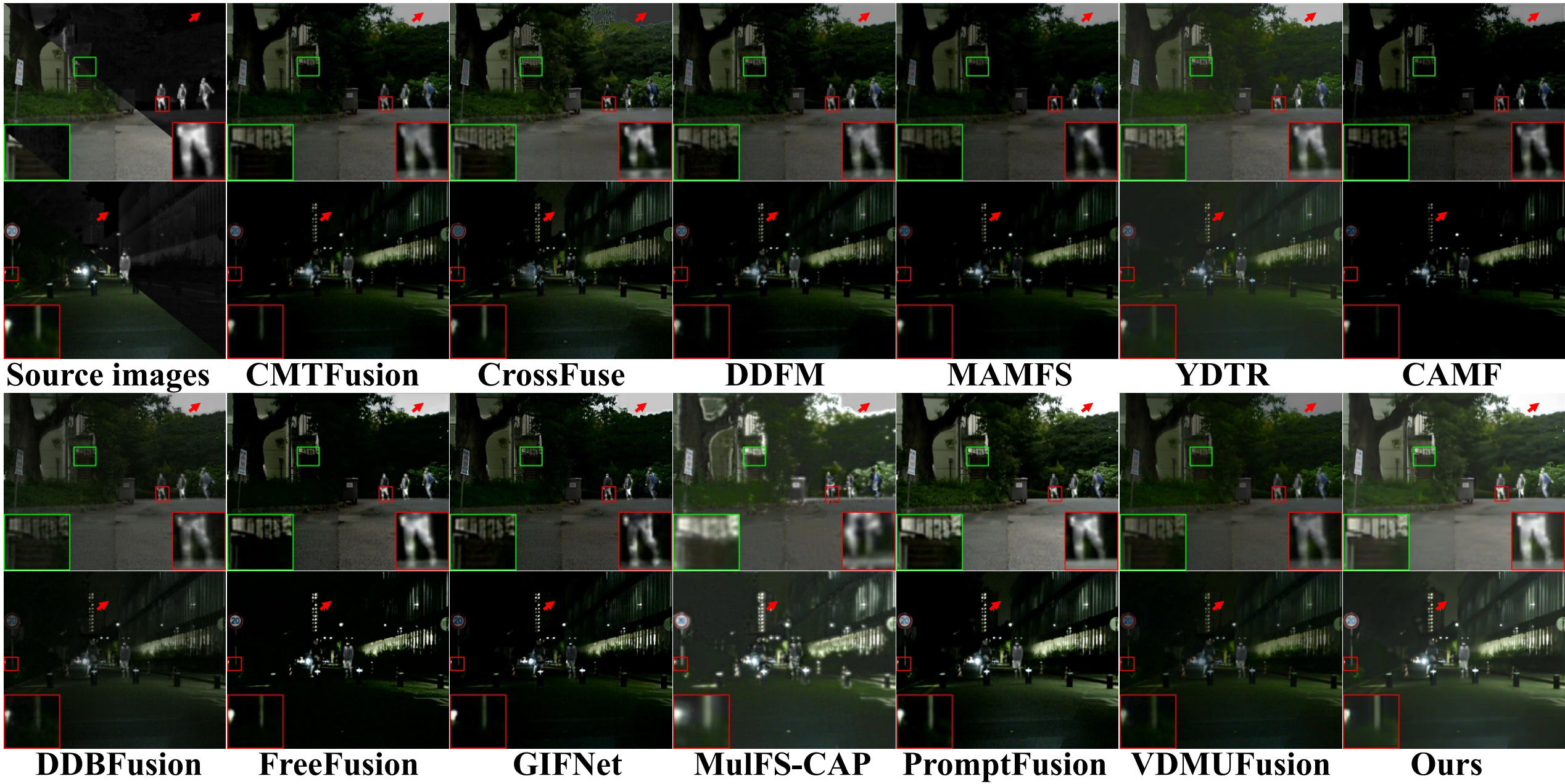


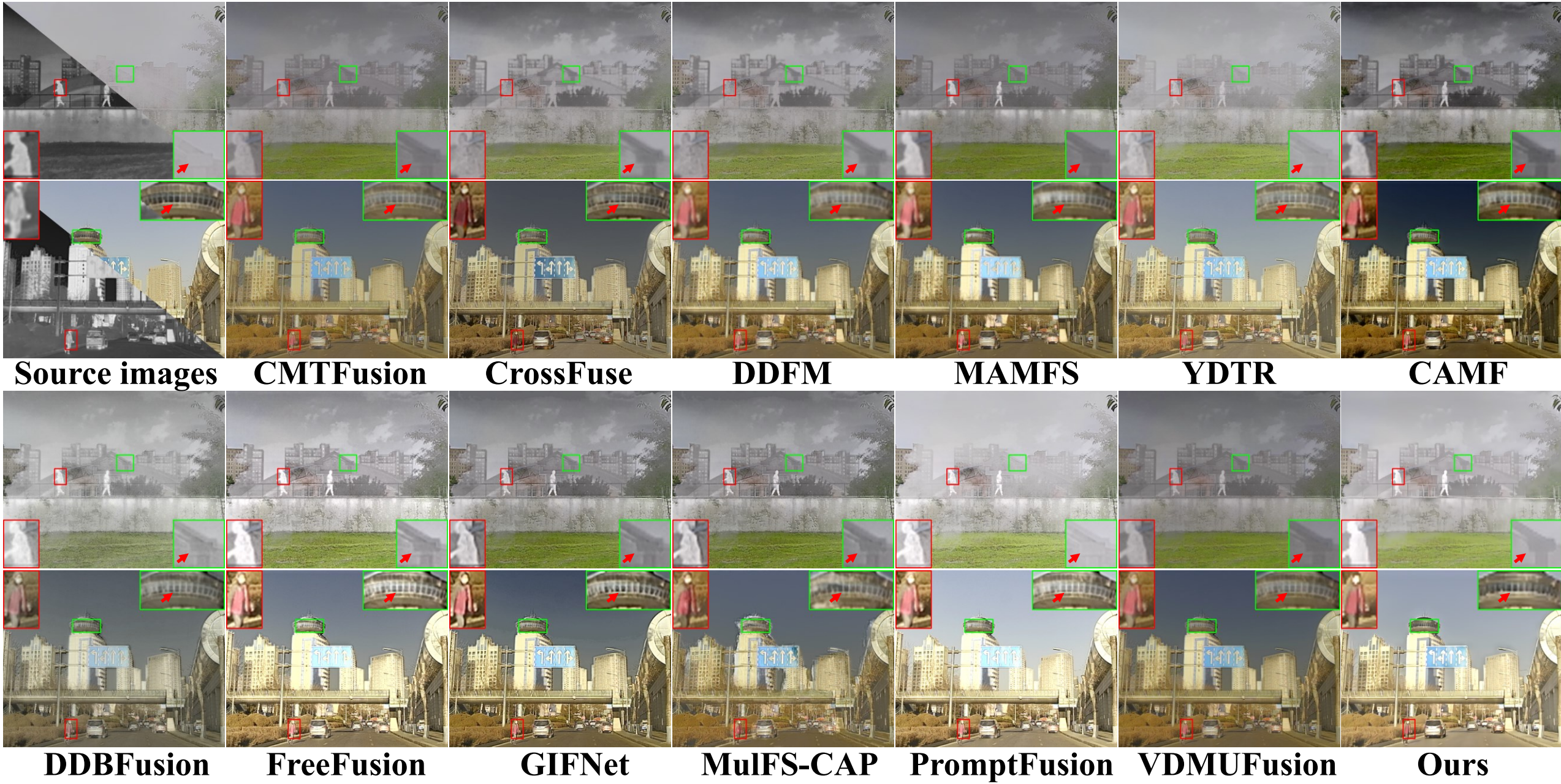
(a) Traditional Feature Fusion Mechanism

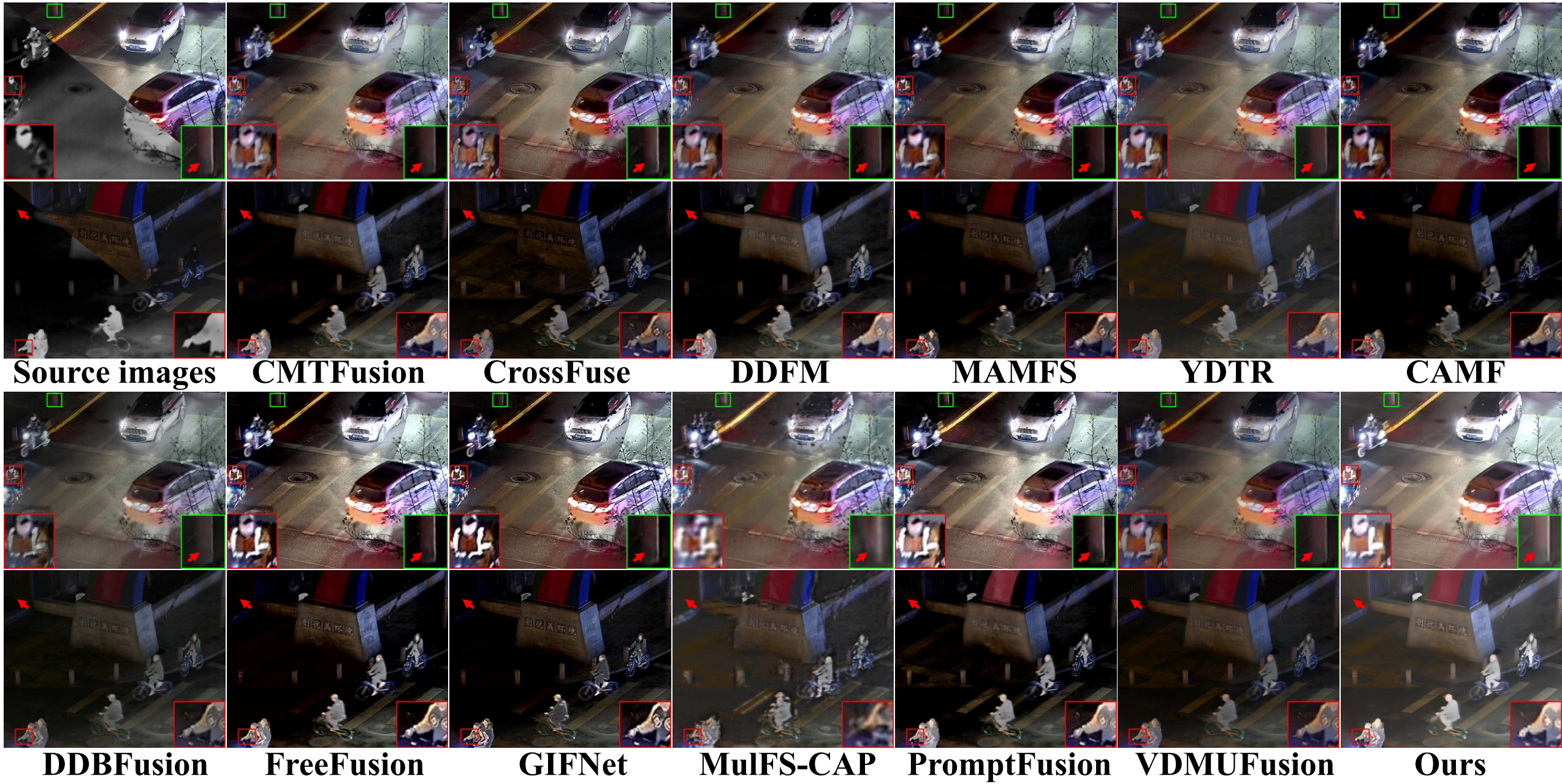
(b) Ours

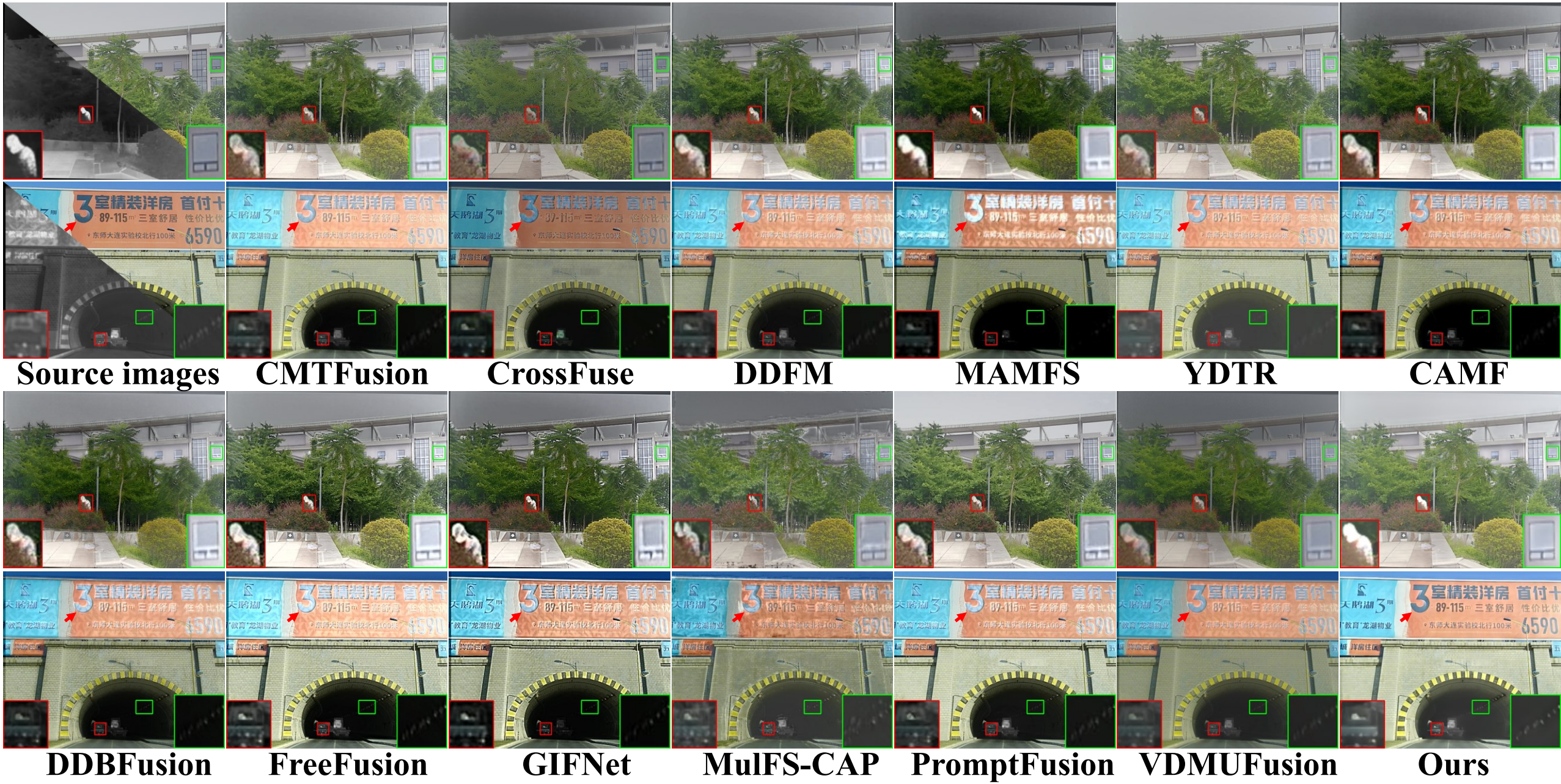


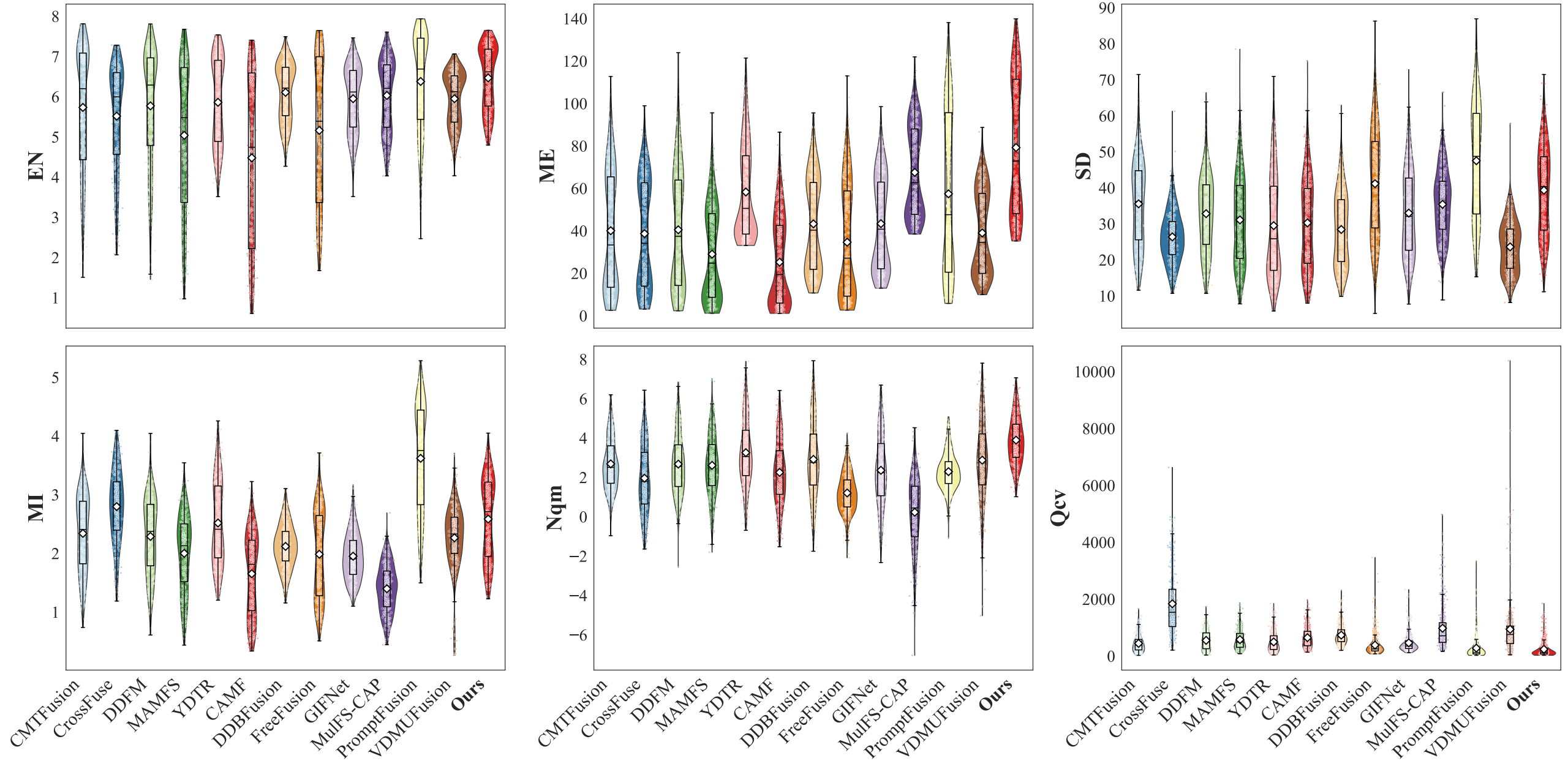
(a) Traditional Loss **(b) VLM based General Loss** **(c) Ours**

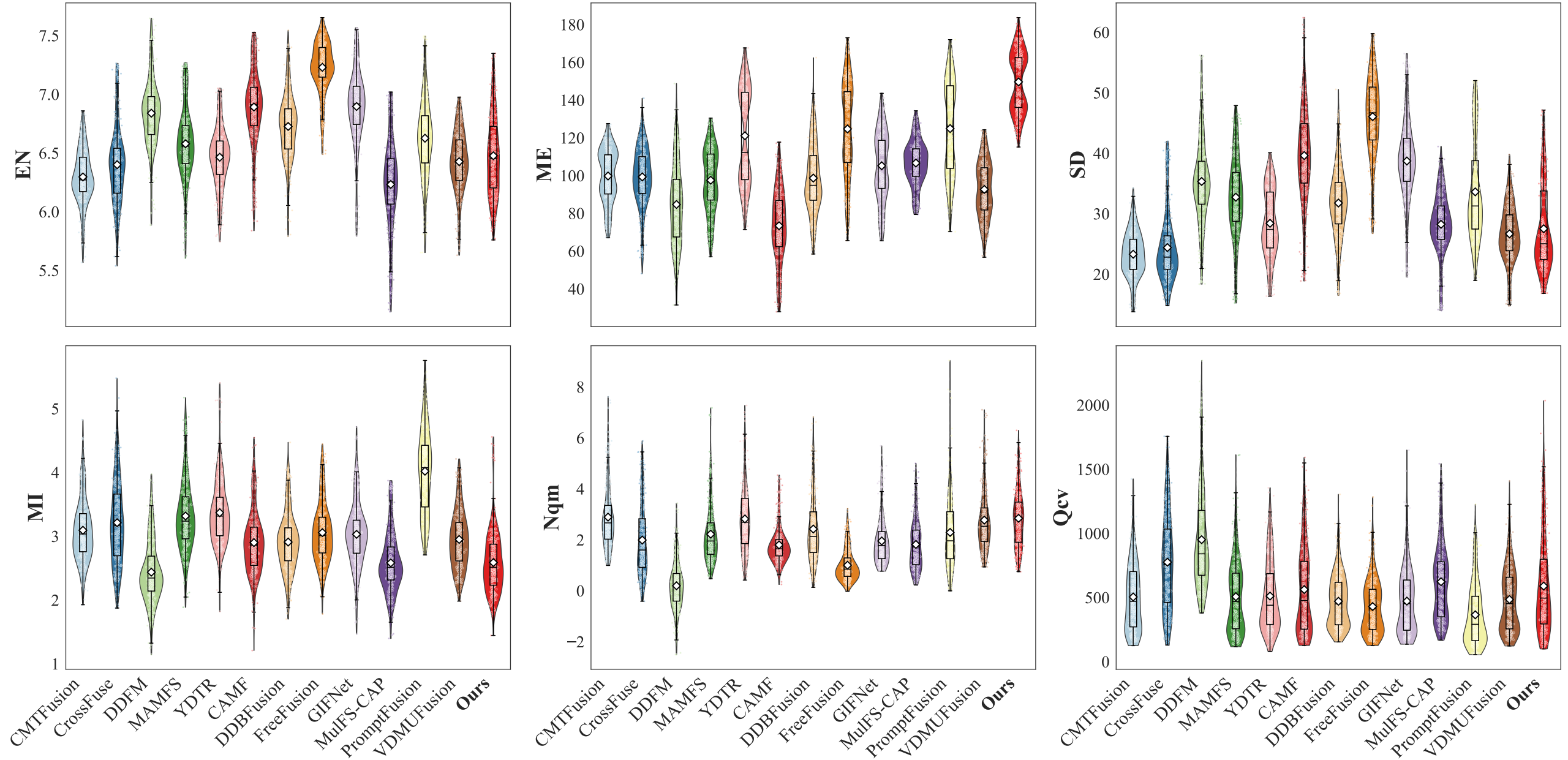

Source images
CMTFusion
CrossFuse
DDFM
MAMFS
YDTR
CAMF
DDBFusion
FreeFusion
GIFNet
MulFS-CAP
PromptFusion
VDMUFusion
Ours

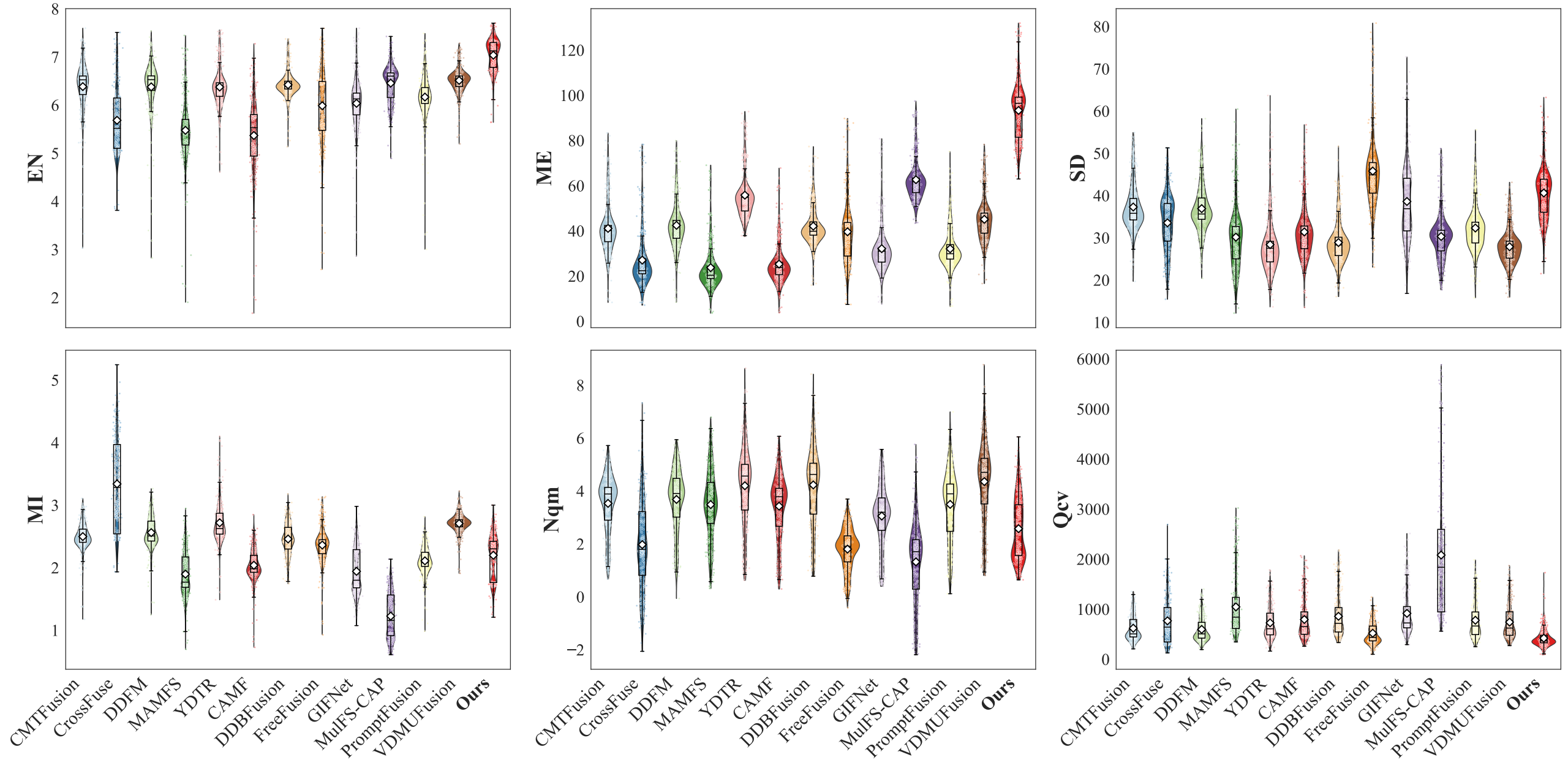

Source images
CMTFusion
CrossFuse
DDFM
MAMFS
YDTR
CAMF
DDBFusion
FreeFusion
GIFNet
MulFS-CAP
PromptFusion
VDMUFusion
Ours

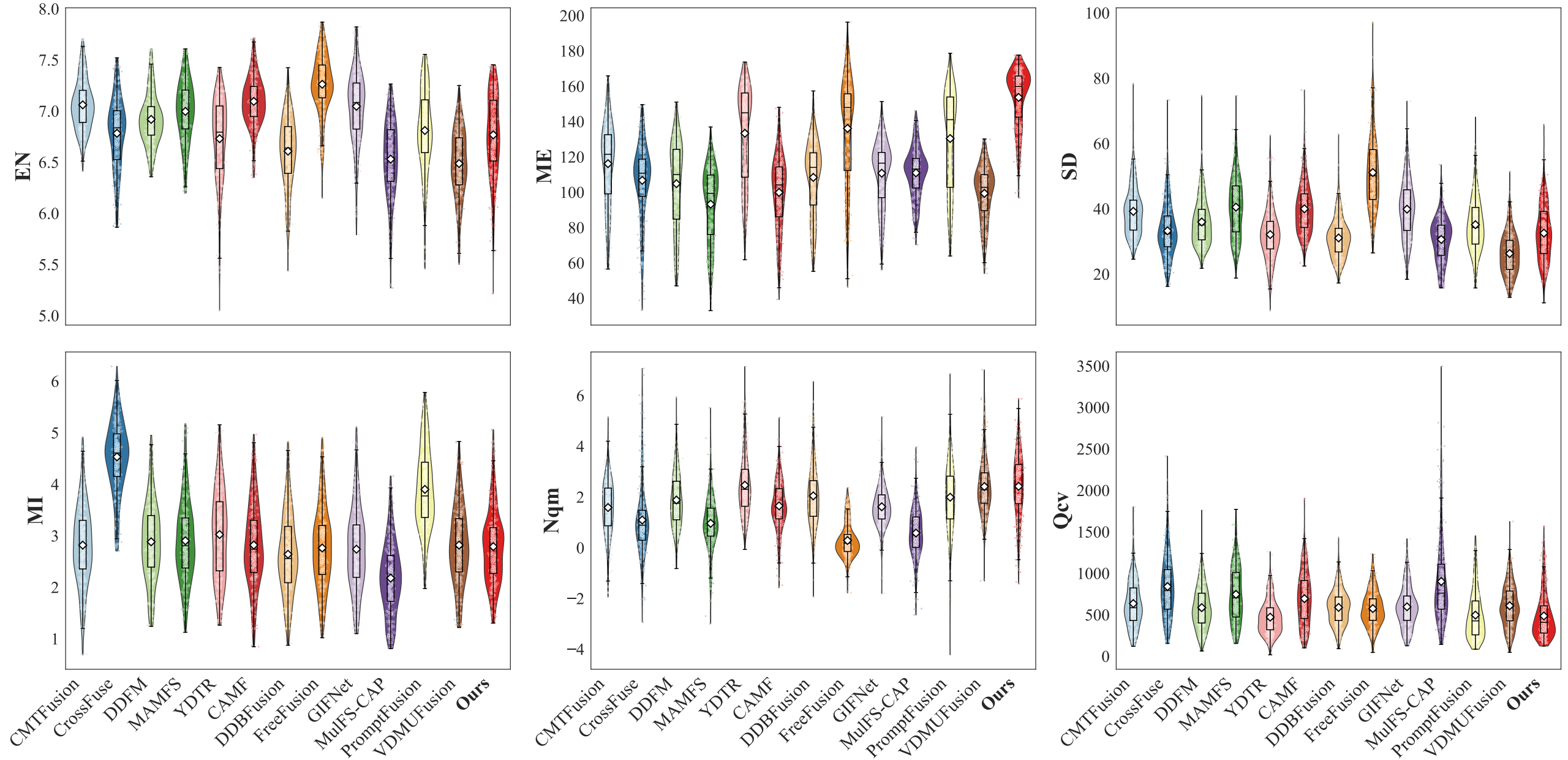

Source images
CMTFusion
CrossFuse
DDFM
MAMFS
YDTR
CAMF
DDBFusion
FreeFusion
GIFNet
MulFS-CAP
PromptFusion
VDMUFusion
Ours

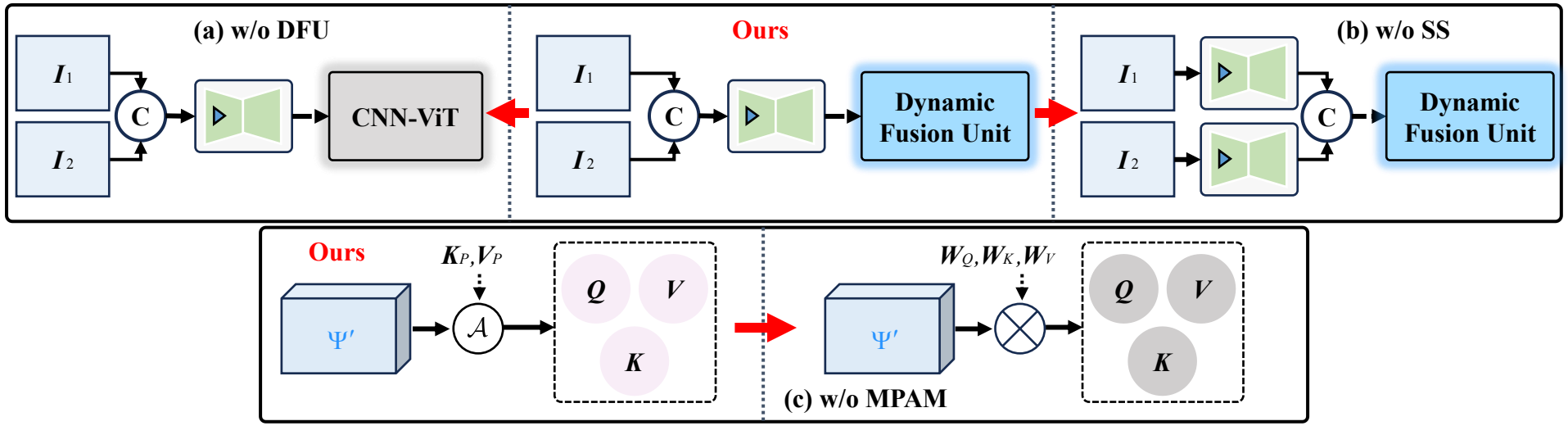

Source images
CMTFusion
CrossFuse
DDFM
MAMFS
YDTR
CAMF
DDBFusion
FreeFusion
GIFNet
MulFS-CAP
PromptFusion
VDMUFusion
Ours


EN
ME
SD
MI
Nqm
Qcv
CMTFusion
CrossFuse
DDFM
MAMFS
YDTR
CAMF
DDBFusion
FreeFusion
GIFNet
MulFS-CAP
PromptFusion
VDMUFusion
Ours


EN
ME
SD
MI
Nqm
Qcv
CMTFusion
CrossFuse
DDFM
MAMFS
YDTR
CAMF
DDBFusion
FreeFusion
GIFNet
MulFS-CAP
PromptFusion
VDMUFusion
Ours


EN
ME
SD
MI
Nqm
Qcv
CMTFusion
CrossFuse
DDFM
MAMFS
YDTR
CAMF
DDBFusion
FreeFusion
GIFNet
MulFS-CAP
PromptFusion
VDMUFusion
Ours

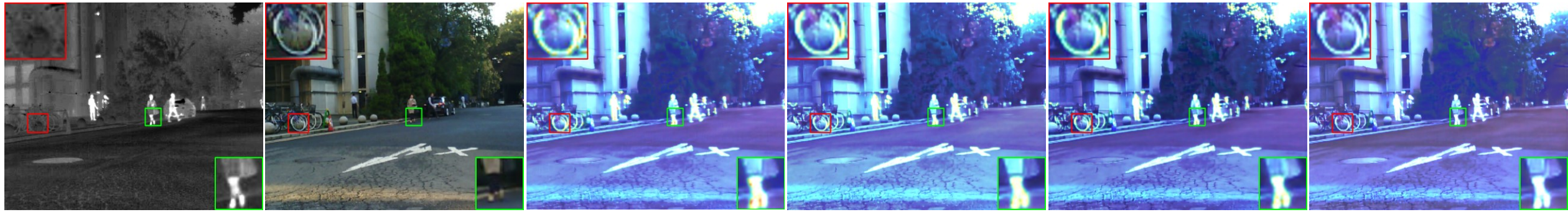

EN
8.0
7.5
7.0
6.5
6.0
5.5
5.0
ME
200
180
160
140
120
100
80
60
40
SD
100
80
60
40
20
MI
6
5
4
3
2
1
Nqm
6
4
2
0
−2
−4
Qcv
3500
3000
2500
2000
1500
1000
500
0
CMTFusion
CrossFuse
DDFM
MAMFS
YDTR
CAMF
DDBFusion
FreeFusion
GIFNet
MulFS-CAP
PromptFusion
VDMUFusion
Ours


(a) w/o DFU
I_1
I_2
C
CNN-ViT
Ours
I_1
I_2
C
Dynamic Fusion Unit
(b) w/o SS
I_1
I_2
C
Dynamic Fusion Unit
Ours
K_P, V_P
Ψ′
$\mathcal{A}$
Q
V
K
W_Q, W_K, W_V
Ψ′
Q
V
K
(c) w/o MPAM

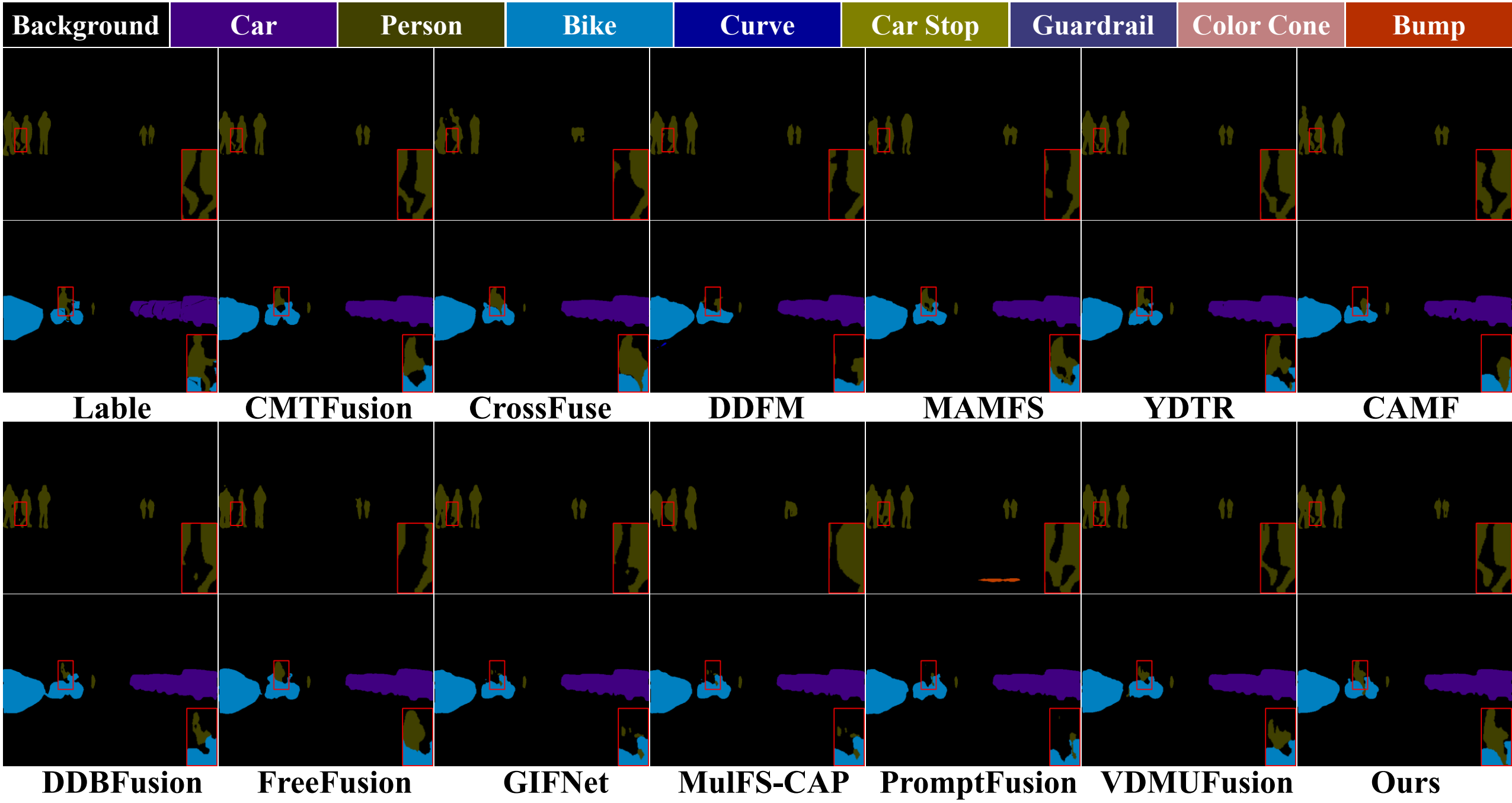
(a) Infrared image (b) Visible image (c) w/o DFU (d) w/o SS (e) w/o MPAM (f) Ours

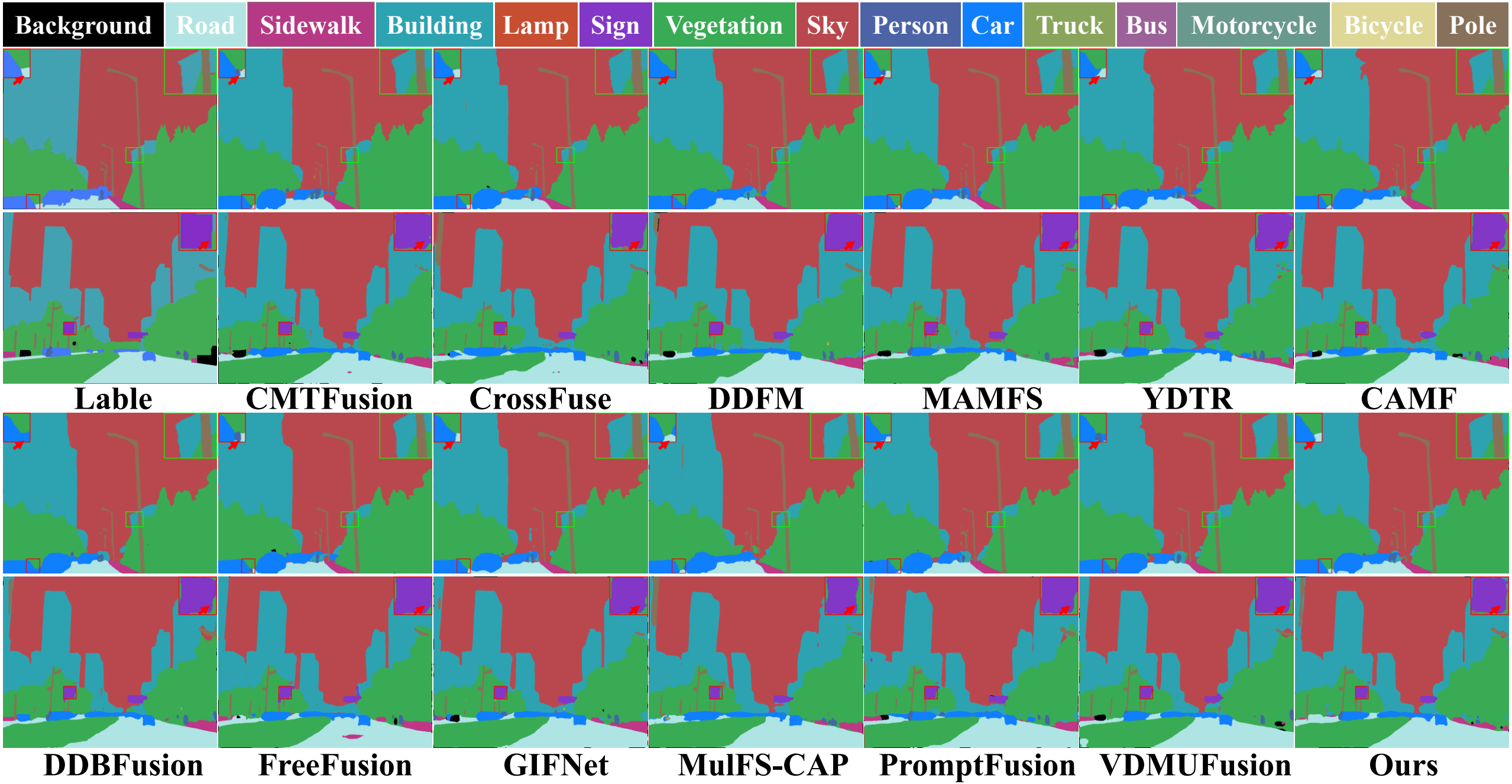
(a) Infrared image (b) Visible image (c) w/o $\mathcal{L}_s$ (d) w/o $\mathcal{L}_p$

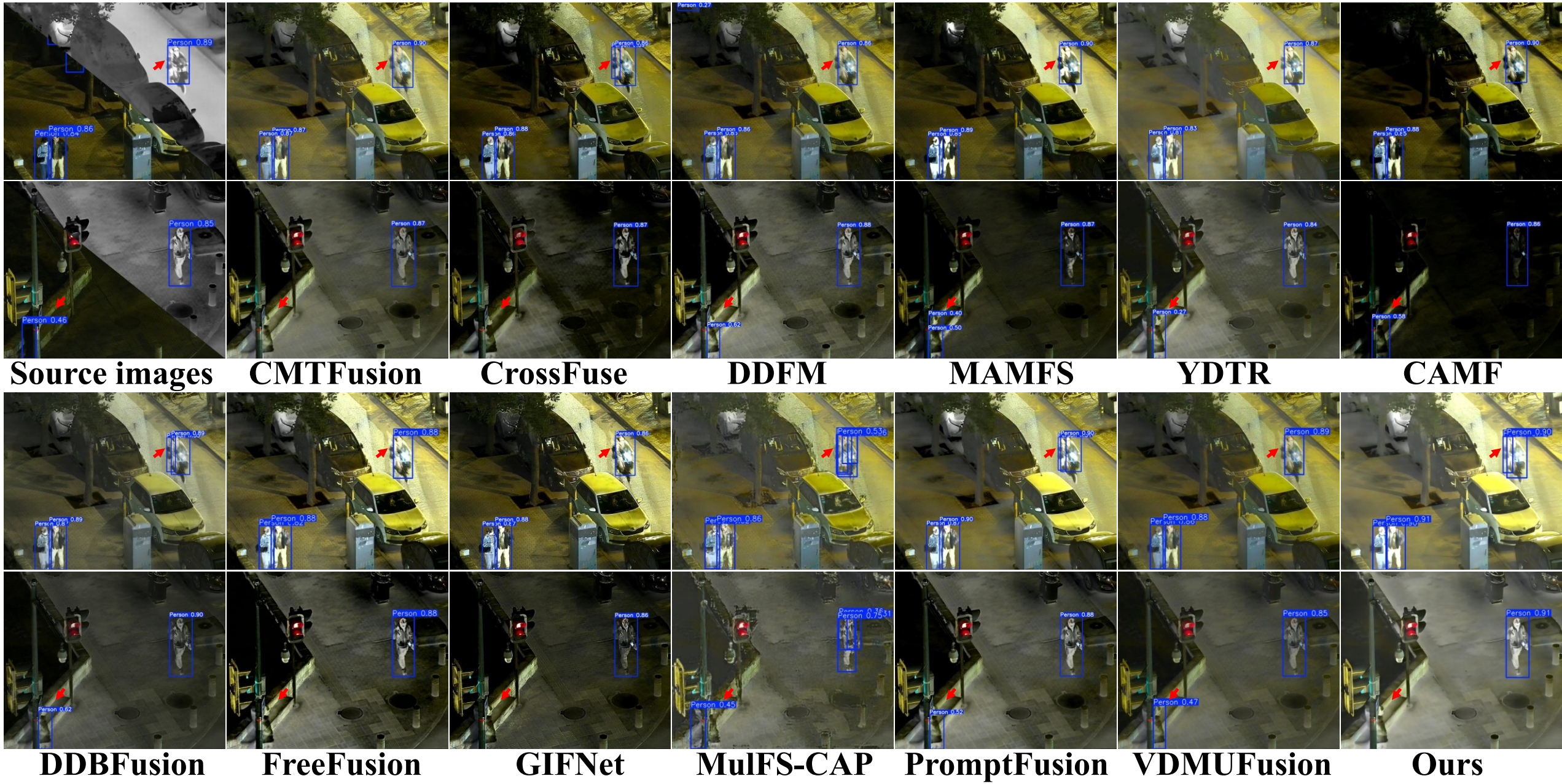
(e) w/o $\mathcal{L}_k$ (f) $\mathcal{L}_k \rightarrow \mathcal{L}_{VLM1}$ (g) $\mathcal{L}_k \rightarrow \mathcal{L}_{VLM2}$ (h) Ours

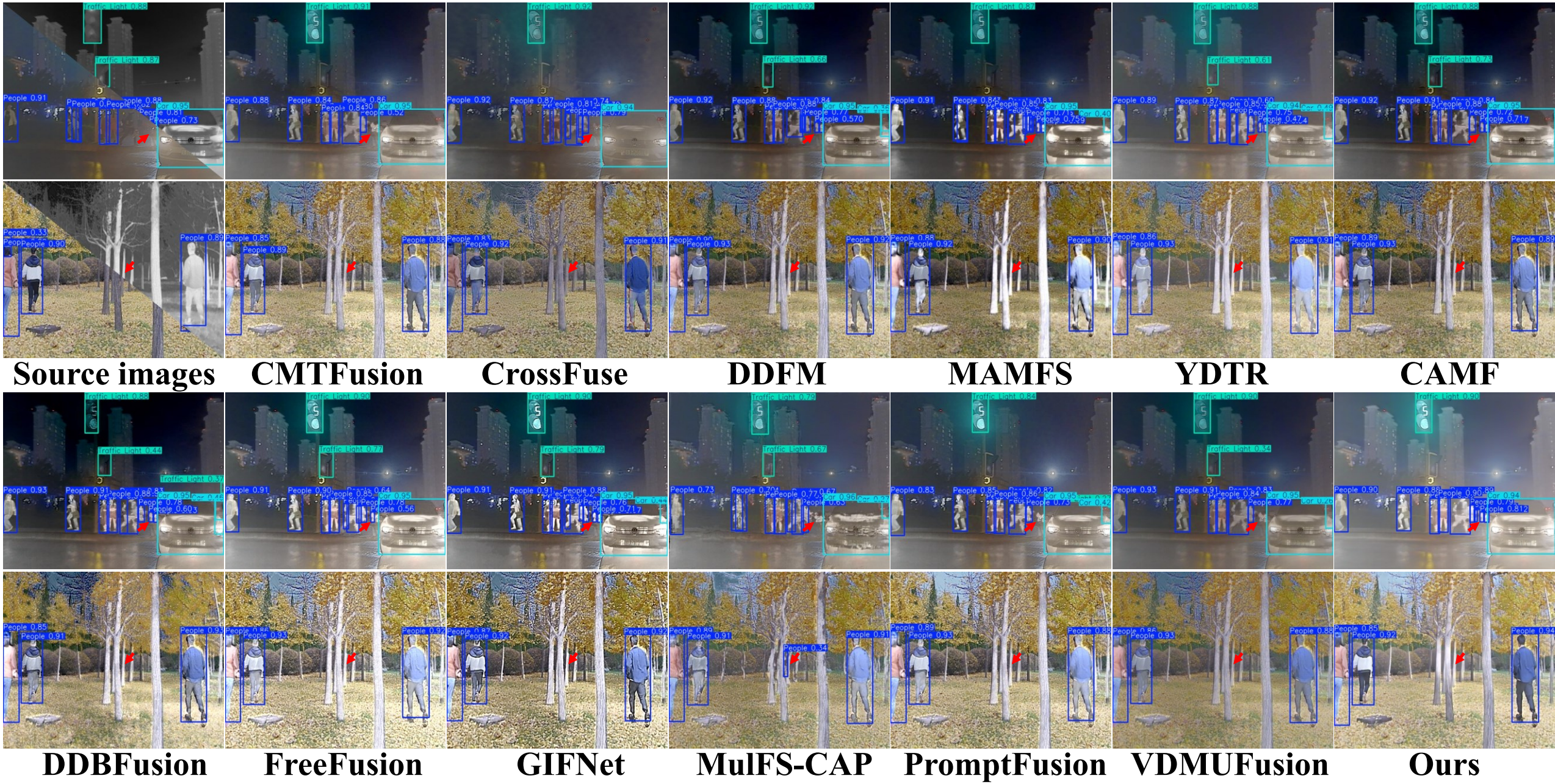


(a) Infrared image (b) Visible image (c) NP (d) CSP (e) GFP (f) Ours


Background
Car
Person
Bike
Curve
Car Stop
Guardrail
Color Cone
Bump
Lable
CMTFusion
CrossFuse
DDFM
MAMFS
YDTR
CAMF
DDBFusion
FreeFusion
GIFNet
MulFS-CAP
PromptFusion
VDMUFusion
Ours


Background
Road
Sidewalk
Building
Lamp
Sign
Vegetation
Sky
Person
Car
Truck
Bus
Motorcycle
Bicycle
Pole
Lable
CMTFusion
CrossFuse
DDFM
MAMFS
YDTR
CAMF
DDBFusion
FreeFusion
GIFNet
MulFS-CAP
PromptFusion
VDMUFusion
Ours


Source images
CMTFusion
CrossFuse
DDFM
MAMFS
YDTR
CAMF
DDBFusion
FreeFusion
GIFNet
MulFS-CAP
PromptFusion
VDMUFusion
Ours


Source images
CMTFusion
CrossFuse
DDFM
MAMFS
YDTR
CAMF
DDBFusion
FreeFusion
GIFNet
MulFS-CAP
PromptFusion
VDMUFusion
Ours